

# Exploratory visualizations and statistical analysis of large, heterogeneous epigenetic datasets

**Author**

**Konstantin Halachev**

**Dissertation**

zur Erlangung des Grades  
des Doktors der Naturwissenschaften (Dr. rer. nat.)  
der Naturwissenschaftlich–Technischen Fakultäten  
der Universität des Saarlandes

Saarbrücken

2014

Tag des Kolloquiums:	10.06.2014
Dekan:	Prof. Dr. Mark Groves
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Dr. h.c. mult. Kurt Mehlhorn
Berichterstatter:	Prof. Dr. Thomas Lengauer, Ph.D. Prof. Dr. Christoph Bock
Beisitzer:	Dr. Nico Pfeifer

## Abstract

Epigenetic marks, such as DNA methylation and histone modifications, are important regulatory mechanisms that allow a single genomic sequence to give rise to a complex multicellular organism. When studying mechanisms of epigenetic regulation, the analyses depend on the experimental technologies and the available data. Recent advancements in sequencing technologies allow for the efficient extraction of genome-wide maps of epigenetic marks. A number of large-scale mapping projects, such as ENCODE and IHEC, intensively produce data for different tissues and cell cultures. The increasing quantity of data highlights a major bottleneck in bioinformatic research, namely the lack of bioinformatic tools for analyzing these data. To date, there are bioinformatics tools for detailed (mostly visual) inspection of single genomic loci, allowing biologists to focus research on regions of interest. Also, efficient tools for manipulation and analysis of the data have been published, but often they require computer science abilities. Furthermore, the available tools provide solutions to only already well formulated biological questions. What is missing, in our opinion, are tools (or pipelines of tools) to explore the data interactively, in a process that would facilitate a trained biologist to recognize interesting aspects and pursue them further until concrete hypotheses are formulated. A possible solution stems from the best practices in the fields of information retrieval and exploratory search. In this thesis, I propose EpiExplorer, a paradigm for integration of state-of-the-art information retrieval methods and indexing structures, applied to offer instant interactive exploration of large epigenetic datasets. The algorithms we use are developed for semi-structured text data, but we apply them on bioinformatic data through clever textual mapping of biological properties. We demonstrate the power of EpiExplorer in a series of studies that address interesting biological problems. We also present in this manuscript EpiGRAPH, a bioinformatic software that we developed with colleagues. EpiGRAPH helps identify and model significant biological associations among epigenetic and genetic properties for sets of regions. Using EpiExplorer and EpiGRAPH, independently or in a pipeline, provides the bioinformatic community with access to large databases of annotations, allows for exploratory visualizations or statistical analysis and facilitates reproduction and sharing of results.



## Kurzfassung

Epigenetische Signaturen wie die Methylierung der DNS oder posttranslationale Modifikationen der Histonproteine stellen wichtige regulatorische Mechanismen dar. Diese ermöglichen es, dass ein komplexer, multizellulärer Organismus aus einer einzelnen genomischen Sequenz hervorgeht. Adequate Analysemethoden hängen von den verwendeten experimentellen Technologien und den verfügbaren Daten ab. Jüngste Fortschritte in der DNS-Sequenzierungstechnologie ermöglichen die effiziente Erstellung genomweiter Karten epigenetischer Informationen. Diese Epigenomkarten werden von einigen Projekten und Initiativen wie ENCODE und IHEC im grossen Massstab für diverse Gewebe- und Zelltypen erstellt. Hierbei stellt der Mangel an effizienten bioinformatischen Softwarewerkzeugen einen wesentlichen Engpass in der Analyse dieser stetig wachsenden Datenflut dar. Experimentelle Biologen können heute einzelne genomische Loci mithilfe benutzerfreundlicher (meist visueller) bioinformatischer Software im Detail inspizieren. Des Weiteren existieren effiziente Werkzeuge für die Manipulation und Analyse dieser Datensätze, die jedoch ein gewisses Mass informatischer Expertise erfordern und sich zumeist auf die Lösung bereits wohldefinierter biologischer Fragestellungen fokussieren. Unserer Ansicht nach fehlen Werkzeuge und Softwarepipelines mithilfe derer ein Benutzer, der über ein fundiertes Wissen der biologischen Grundlagen, jedoch nicht unbedingt über informatische Kenntnisse verfügt, die verfügbaren Datensätze interaktiv durchstöbern und darauf aufbauend weiterführende Hypothesen entwickeln kann. Eine möglichen Ansatz hierfür bieten Methoden aus den Bereichen Information Retrieval und der explorativen Suche. Diese Arbeit beschreibt EpiExplorer, eine Software, die auf dem Paradigma der Integration von modernen Information Retrieval und Indexstrukturen basiert und darauf ausgelegt ist eine Vielzahl von (epi-)genomweiten Datensätzen in Echtzeit zu explorieren. Die verwendeten Algorithmen wurden ursprünglich für die Suche in semistrukturierten, textuellen Datensätzen entwickelt. EpiExplorer ermöglicht ihre Verwendung durch eine systematische Umwandlung biologischer Eigenschaften in Textdokumente. Ausserdem demonstriert diese Arbeit EpiExplorers Leistungsfähigkeit und Nützlichkeit durch relevante Anwendungsbeispiele biologisch interessanter Fragestellungen. Komplementär zu EpiExplorer wurde in Kollaboration mit Kollegen EpiGRAPH entwickelt, mithilfe dessen signifikante biologische Assoziationen zwischen genetischen und epigenetischen Eigenschaften regionsbasiert identifiziert und modelliert werden können. EpiExplorer und EpiGRAPH stellen - unabhängig voneinander oder im Verbund miteinander - nützliche Ressourcen dar. In einer bioinformatischen Softwarepipeline ermöglichen sie den Datenbank-basierten Zugriff auf eine Vielzahl (epi-)genomischer Datensätze, deren explorative Visualisierung oder statistische Analyse sowie die Reproduzierbarkeit und den Austausch von Analyseergebnissen.



## Acknowledgements

First, I want to thank my supervisors Thomas Lengauer and Christoph Bock. Thank you Thomas, for providing me with continuous support, encouragement and good advice through the years. Thank you Christoph, for introducing me to the interesting field of epigenetics and sharing numerous discussions on biology, epigenetics, software engineering and statistical methods.

Special thanks to all colleagues with whom I worked at the Department for Computational Biology and Applied Algorithms at the *Max Planck Institute for Informatics*, especially to Lars Feuerbach, Fabian Müller, Felipe Albrecht, Peter Ebert, Lars Steinbrück, Sven-Eric Schelhorn, Oliver Sander, Adrian Alexa and Fidel Ramirez.

I am grateful to Joachim Büch and Georg Friedrich for assisting with my various technical requests. I am much obliged to Ruth Schnepfen-Christmann for shielding me from multiple bureaucratic tasks through the years.

I want to say thanks to the Bulgarian community in Saarbrücken that I hope to meet again at the usual time of the week at the football field, or at least at the Kleine Tonhalle.

I want to thank my colleagues at SDL Fredhopper, especially Pavel Penchev and Nikolay Diakov, for their interest and support for this thesis.

I appreciate the patience of my very close friends Hristo Ganev, Stefan Kiryakov and Pavlin Nedelchev who often listened to monologues on the topic of epigenetic modifications, sometimes into the small hours.

Finally, I want to thank a number of special friends: Yassen Assenov, Rayna Dimitrova, Evangelia Pyrga, Andre Altmann, Dimitar Denev, Jasmina Bogojenska and Hagen Blankenburg. Thank you for all the great moments and hope to have many more.

I want to express my gratitude my parents Iskra, Anastas and my brother Lyubomir for their love and support. And Elena, for teaching me that it is ok to sleep only 4 hours a day.

Ultimately, I want to thank Laura, for being there for me, every day. Thank you Lau, for everything!





# Contents

<b>1. Introduction</b>	<b>7</b>
1.1. Overview . . . . .	8
<b>2. Computational epigenetics and text retrieval</b>	<b>9</b>
2.1. The genome. Genomic annotations. . . . .	9
2.2. Epigenetics . . . . .	12
2.3. DNA methylation . . . . .	13
2.3.1. Functions and mechanisms . . . . .	13
2.3.2. The dynamics of DNA methylation during embryonic development and tissue differentiation . . . . .	14
2.3.3. Experimental technologies . . . . .	17
2.4. Histone modifications . . . . .	19
2.4.1. Dynamic regulation via histone modifications . . . . .	19
2.4.2. Experimental technologies . . . . .	20
2.5. Computational analysis of genomic and epigenomic data . . . . .	22
2.5.1. Overview of bioinformatic tools for working with epigenome annota- tions . . . . .	22
2.5.2. The characteristics of existing bioinformatic tools . . . . .	23
2.6. Text retrieval and exploratory search . . . . .	25
2.6.1. Information retrieval and text retrieval . . . . .	25
2.6.2. Data indexing and query types . . . . .	26
<b>3. Live exploration and global analysis of large epigenomic datasets using EpiEx- plorer</b>	<b>29</b>
3.1. Background . . . . .	30
3.2. Concept and main features of EpiExplorer . . . . .	31
3.2.1. Functionalities of the EpiExplorer software server . . . . .	33
3.3. Methods . . . . .	36
3.3.1. Translating biological concepts to text . . . . .	36
3.3.2. Precomputing epigenomic and genomic properties of genomic regions	39
3.3.3. Software architecture . . . . .	42
3.3.4. Computation workflow when processing a typical EpiExplorer query	45
3.3.5. EpiExplorer user interface and user experience . . . . .	47
3.4. Applications of EpiExplorer . . . . .	54
3.4.1. Rediscovering properties of CpG islands and application for discovery of robust CpG island annotations . . . . .	54
3.4.2. Connecting a new epigenetic mark to reference maps of the human genome and epigenome . . . . .	61

3.4.3. Epigenetic properties of cancer breakpoints . . . . .	69
3.4.4. EpiExplorer performance evaluation . . . . .	73
3.4.5. EpiExplorer usage statistics . . . . .	74
3.5. Conclusions and outlook . . . . .	77
<b>4. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data</b>	<b>81</b>
4.1. Background . . . . .	82
4.1.1. EpiGRAPH overview . . . . .	82
4.2. Methods . . . . .	84
4.2.1. Mapping of genomic and epigenomic annotations . . . . .	85
4.2.2. Statistical and machine learning analyses of EpiGRAPH . . . . .	89
4.2.3. Software architecture and implementation of the EpiGRAPH service	91
4.3. Applications . . . . .	95
4.3.1. Running example: DNA methylation of CpG islands . . . . .	95
4.3.2. DNA methylation in pluripotent cells may constitute an epigenetic ground state . . . . .	101
4.3.3. Epigenetics of orthologous gene promoters . . . . .	110
4.4. Conclusions . . . . .	118
<b>5. Conclusions</b>	<b>119</b>
5.1. Outlook . . . . .	120
<b>Appendices</b>	<b>121</b>
<b>A. EpiExplorer annotations listing</b>	<b>123</b>
A.1. Human genome . . . . .	123
A.1.1. hg19 . . . . .	123
A.1.2. hg18 . . . . .	132
A.2. Mouse genome . . . . .	139
A.2.1. mm9 . . . . .	139
<b>B. EpiGRAPH attribute reference sheet</b>	<b>143</b>
B.1. Overview . . . . .	143
B.2. DNA sequence attributes (calculated from pattern frequencies) . . . . .	143
B.3. DNA structure attributes (calculated from oligomers with known structure)	144
B.4. Patch attributes (quantifying overlap with sets of genomic regions) . . . . .	144
B.5. Gene attributes (quantifying overlap with genes and exons) . . . . .	146
<b>Bibliography</b>	<b>147</b>

# List of Figures

2.1. Structure of a gene and the steps to transcription . . . . .	10
2.2. Representation of DNA methylation and histone modifications, the two main actors in epigenetic regulation . . . . .	12
2.3. Two reprogramming stages of DNA methylation . . . . .	15
2.4. DNA methylation deposition and maintenance by DNA methyltransferases .	16
2.5. Schematic representation of bisulfite sequencing . . . . .	18
2.6. Schematic representation of chromatin immunoprecipitation . . . . .	21
3.1. Schematic outline of genomic regions uploaded in a BED format. . . . .	39
3.2. Schematic outline of genomic regions annotated with multiple genomic and epigenomic annotation properties . . . . .	40
3.3. Genomic regions are converted into text documents and each property is represented as a word in the document. . . . .	40
3.4. To create the HYB index, CompleteSearch needs the sorted lists of all words and document identifiers. . . . .	41
3.5. Schematic representation of the HYB index structure. . . . .	41
3.6. Schematic outline of EpiExplorer's software architecture, consisting of a web-based user interface, a query-processing and annotation-mapping middleware, and a text-search backend. . . . .	43
3.7. Overview of how EpiExplorer processes a set of genomic regions . . . . .	45
3.8. Illustration how EpiExplorer processes the request to retrieve all regions overlapping with CpG islands from a sample dataset . . . . .	46
3.9. Illustration how EpiExplorer processes the request to retrieve all regions overlapping with CpG islands and an H3K4me3 peak . . . . .	46
3.10. Illustration how EpiExplorer requests and provides faceting information for a set of genomic regions with a single prefix query . . . . .	47
3.11. Selecting a dataset view of EpiExplorer . . . . .	47
3.12. Exploring a dataset view in EpiExplorer . . . . .	48
3.13. Refining a selection of an EpiExplorer dataset . . . . .	49
3.14. Uploading a custom dataset . . . . .	51
3.15. Dynamic status when preprocessing a custom dataset . . . . .	51
3.16. The compare button (highlighted) activates the mode that allows direct comparison of two region selections . . . . .	53
3.17. At all times in comparison mode, the interface lists the current selection as well as the full details of the reference dataset . . . . .	53
3.18. Select the unlock button to activate the dynamic reference mode . . . . .	54
3.19. Bar chart summarizing the percent overlap (y-axis) between CpG islands and various genomic region sets (x-axis) in H1hESC cells. . . . .	55

3.20. Bubble chart plotting the percent overlap (y-axis) between CpG islands and H3K4me3 peaks in specific tissues (color-coded) against the total genomic coverage of all corresponding peaks (x-axis) . . . . .	56
3.21. Neighborhood plot illustrating the percent overlap (y-axis) with histone H3K4me3 peaks in the vicinity of CpG islands (x-axis). Line colors correspond to histone modification data for different cell types. . . . .	57
3.22. Neighborhood plot illustrating the percent overlap (y-axis) with histone H3K27me3 peaks in the vicinity of CpG islands (x-axis). Line colors correspond to histone modification data for different cell types. . . . .	57
3.23. Percent overlap (y-axis) of 13,519 CpG islands located within one kilobase from a gene transcription start site (orange) and 2,327 CpG islands located at least 20 kilobases from the nearest gene (gray) with genome and epigenome annotation data (x-axis) . . . . .	58
3.24. Percent overlap (y-axis) of 15,377 constitutively unmethylated CpG islands (orange, less than 30% methylation in seven tissues) and 3,171 constitutively methylated CpG islands (grey, more than 60% methylation in the same seven tissues) with genome and epigenome annotation data (x-axis). . . . .	58
3.25. Overview of the length distribution of constitutively unmethylated CpG islands (left) and constitutively methylated CpG islands (right). . . . .	59
3.26. Distribution of CpG dinucleotide frequencies among constitutively unmethylated CpG islands (orange) and among constitutively methylated CpG islands (grey). . . . .	60
3.27. Distribution of TpG dinucleotide frequencies among constitutively unmethylated CpG islands (orange) and among constitutively methylated CpG islands (grey). . . . .	60
3.28. Bar chart summarizing the percent overlap (y-axis) between 5hmC hotspots and various genomic datasets (x-axis) in H1hESC cells. . . . .	61
3.29. Bar chart comparing the percent overlap of 5hmC hotspots (orange) and randomized control regions (grey) with histone H3K4me1 peaks, based on ENCODE data (Myers et al., 2011). . . . .	62
3.30. Genomic neighborhood plot illustrating the percent overlap (y-axis) with H3K4me1 peaks in the vicinity of 5hmC hotspots (x-axis). Different line colors correspond to H3K4me1 data for different cell types. . . . .	63
3.31. Bar chart comparing the percent overlap of 5hmC hotspots (orange) and randomized control regions (grey) with a comprehensive catalog of epigenetic states derived by computational segmentation of ENCODE histone modification data (Ernst et al., 2011). . . . .	63
3.32. Distribution of DNA methylation levels among 5hmC hotspots (orange) and randomized control regions (grey), based on Roadmap Epigenomics data (Human Epigenome Atlas, 2013). . . . .	64
3.33. Enrichment table (left) and word cloud (right) illustrating the most highly enriched Gene Ontology (GO) terms among genes whose transcribed region is within 10 kb of a 5hmC hotspot. The most general (more than 5,000 associated genes) and most specific GO terms (less than 50 associated genes) were suppressed in this analysis. . . . .	64

3.34. Using successive filtering steps, a genomic dataset with 82,221 hotspots of 5-hydroxymethylcytosine (5hmC) in human ES cells (Szulwach et al., 2011) is refined to a list of 16 regions that provide strong candidates for investigating the functional association between 5hmC and H3K4me1-marked enhancer elements. (a) Filtering with a minimum length threshold of 1 kb yields 5,734 genomic regions. . . . .	66
3.35. Filtering with a minimum 5hmC hotspot score threshold of 300, which corresponds to a detection significance of 10-30 or better, yields 2,535 genomic regions. . . . .	66
3.36. Filtering for overlap with H3K4me1 peaks in a human ES cell line (H1hESC) yields 2,334 genomic regions. . . . .	67
3.37. Filtering for association with genes that are annotated with any of the 1,608 Gene Ontology terms containing the word 'regulation' yields 1,064 genomic regions. . . . .	67
3.38. Filtering for overlap with an alternative dataset of 5hmC hotspots (Stroud et al., 2011) yields 99 genomic regions. . . . .	68
3.39. Filtering for a minimum DNA methylation coverage threshold of five CpGs yields 65 genomic regions. . . . .	68
3.40. Filtering for intermediate DNA methylation with levels in the range of 20% to 50% yields 16 genomic regions. . . . .	68
3.41. EpiExplorer screenshot showing the final list of candidate regions, ready for visualization in a genome browser, for download and manual inspection, and for export to other web-based tools for further analysis. . . . .	69
3.42. Overlap of breakpoints from seven different cancer types with CpG islands .	71
3.43. Overlap of breakpoints from seven different cancer types with H3K3me3 peaks in ES cells . . . . .	71
3.44. Overlap of breakpoints from seven different cancer types with H3K27me3 peaks in ES cells . . . . .	72
3.45. Overlap of breakpoints from seven different cancer types with insulators in ES cells . . . . .	72
3.46. Number of EpiExplorer analysis per month in its first year . . . . .	75
3.47. Number of EpiExplorer analysis per genome per month in its first year . .	75
3.48. Distribution of performed EpiExplorer analyses for different epigenetic annotations . . . . .	76
3.49. Number of computed custom datasets per month during EpiExplorer's first year . . . . .	77
4.1. EpiGRAPH analysis workflow. . . . .	84
4.2. Overlap of a region with a patch annotation . . . . .	86
4.3. Overlap of a region with a patch annotation with scores . . . . .	87
4.4. EpiGRAPH software architecture . . . . .	92
4.5. Results of EpiGRAPH statistical module for CGIs with constitutive methylation . . . . .	97
4.6. Results of EpiGRAPH statistical module for Yamada et al. (2006) dataset .	98
4.7. Diagram visualizing the difference of the CpA/TpG distributions . . . . .	99

4.8. Machine learning analysis results when predicting methylation of CpG islands with constitutive methylation . . . . .	99
4.9. Machine learning analysis results when predicting methylation of CpG islands from Yamada et al. (2006) . . . . .	100
4.10. Performance of linear SVM models trained with combinations of groups of features (rows). . . . .	100
4.11. Performance of various machine learning models, trained with DNA sequence features. . . . .	101
4.12. DNA sequence features that were identified to be significantly associated with DNA methylation state of BFCGIs. . . . .	108
4.13. Histogram of DNA sequence patterns that were found significant in at least one tissue. . . . .	109
4.14. Heatmap representation of the prediction accuracies from machine learning classification scenarios for multiple tissues and region types. . . . .	110
4.15. Statistical results of the EpiGRAPH analysis of CGI-state of mouse promoters orthologous to human CGI promoters . . . . .	113
4.16. Box plot showing the distributions of observed versus expected ratios of CpG counts of human CpG island promoters that are orthologous to mouse CGI promoters (yellow) or mouse non-CGI promoters (red). . . . .	114
4.17. Machine learning results of the EpiGRAPH analysis of CGI-state of mouse promoters orthologous to human CGI promoters . . . . .	114
4.18. Visualization of some of the most significant features differentiating between methylated and unmethylated mouse non-CGI promoters orthologous to human CGI promoters. . . . .	116
4.19. List of genomic features that significantly differ between human and mouse promoters. . . . .	117

# List of Tables

2.1. Characteristics of the most popular bioinformatic tools for analysis of epigenome data. . . . .	24
3.1. Number of tumor, samples. breakpoints and consensus breakpoints identified in sever cancer datasets . . . . .	70
3.2. EpiExplorer's response time and memory footprint across thousands of actual user analyses. . . . .	74
3.3. Overview statistics of EpiExplorer's first year. . . . .	74
4.1. Numbers of attributes included in EpiGRAPH by type and genome assembly. . . . .	86
4.2. Contingency table of attributes $X$ and $Y$ for Fisher's exact test. . . . .	90
4.3. Methylation of CpGs in multiple mouse RRBS experiments from Meissner et al. (2008); Gu et al. (2010) sorted by number of methylated CpGs . . . . .	104
4.4. Methylation of CpG islands in multiple mouse RRBS experiments from Meissner et al. (2008); Gu et al. (2010) sorted by number of methylated CGIs . . . . .	105
4.5. Prediction accuracies on the DMR BFCGI methylation in multiple mouse RRBS experiments based on an SVM model trained on the consistently methylated BFCGIs . . . . .	106
4.6. Prediction accuracies of the BFCGI methylation in multiple mouse RRBS experiments based on SVM models trained for each individual tissue . . . . .	107
4.7. Conservation of CpG islands in promoters orthologous in human and mouse . . . . .	111
4.8. Distribution of methylation data for orthologous promoters visualized by genome and promoter CGI status . . . . .	115
A.1. Full listing of the EpiExplorer's annotation datasets for human genome assembly hg19. . . . .	123
A.2. Full listing of the EpiExplorer's annotation datasets for human genome assembly hg18. . . . .	132
A.3. Full listing of the EpiExplorer's annotation datasets for mouse genome assembly mm9. . . . .	139





# 1. Introduction

Since the discovery of the genome, scientists have sought the secrets of the structure, evolution and complexity of mammalian life that it carries. The mammalian genome encodes genes that drive the function of complex organisms with trillions of cells. The genomes of those cells are close to identical as they originate from a single cell (*zygote*). Life has evolved regulatory mechanisms that utilize the identical genome to serve the needs of different cell types. Recent technological advancements reveal intricate functions of *epigenetic* regulatory mechanisms that regulate the access of transcriptional mechanisms to the DNA sequence, are faithfully maintained over cell divisions, but are not encoded in the DNA sequence itself. In this work, we focus on the two best known epigenetic marks, *DNA methylation* and *histone modifications*.

When studying epigenetic mechanisms of regulation, the quality of insights depends highly on the experimental technology and data available. Recent advancements in sequencing technologies (Schuster, 2007) combined with bisulfite treatment (for DNA methylation, Frommer et al. (1992)) and chromatin immunoprecipitation (ChIP) enable the extraction of genome-wide maps of epigenetic regulation. Data have already been collected for different tissues and cell cultures. The ever increasing quantity of data pressures the bioinformatics community to continuously provide tools for visualizing data, extracting meaningful associations and testing statistical hypotheses. To date, the most popular bioinformatic tools, genome browsers, support detailed (mostly visual) inspection of only individual genomic loci (Karolchik et al., 2011; Flicek et al., 2008; Zhou et al., 2011). Efficient tools for analysis of multiple genomic locations in concert exist (Goecks et al., 2010; Quinlan and Hall, 2010), but often provide solutions to only already well formulated biological questions that themselves require deeper understanding of the data and underlying biological processes. What is missing, in my opinion, is a tool (or a pipeline of tools) to explore epigenetic data interactively, in a process that would facilitate a trained biologist to recognize interesting aspects and pursue them further until concrete hypotheses are formulated. A possible inspiration for such methodology could be adapted from state-of-the-art methods in information retrieval, e-commerce and web search engines. Information retrieval methods focus on efficient data indexing that allows quickly answering to a variety of queries about the data. E-commerce applications focus on identifying intuitive ways to present and suggest product data to the users, often assuming that the user is not familiar with the product catalog. Web search engines, used by hundreds of millions of people every day, use information retrieval methods to enable access to large amounts of heterogeneous web data that users query using natural language.

In this thesis, we propose two tools available as public software services called EpiExplorer and EpiGRAPH, and we describe the methodology involved. With EpiExplorer, I prototyped a new paradigm for interactive visual exploration of large genetic and epigenetic data based on concepts from text indexing and information retrieval. EpiExplorer enables

users to interactively inspect via dynamic visualizations the association between custom datasets and a variety of public genetic and epigenetic annotations. The near-instant query responses and the intuitive interface makes the web service a convenient mediator that helps the biologists ask questions, receive answers and interpret them in a simple interactive process that does not require knowledge in programming or statistics. EpiExplorer helps sift through biological data, but it does not provide definitive answers, in a rigorous mathematical or statistical sense. For this goal, one can pipe EpiExplorer analysis into the rigorous statistical framework that EpiGRAPH offers. EpiGRAPH uses statistical and machine learning methods to establish associations between genomic and epigenomic properties of a set of regions. There are large database of epigenetic and genetic annotations underlying both tools that users can easily analyze without facing the challenges of data handling or the necessity of programming skills. In combination, EpiExplorer and EpiGRAPH assist biologists and bioinformaticians with the integration of epigenetic and genetic datasets, extraction of interesting hypotheses and testing their validity.

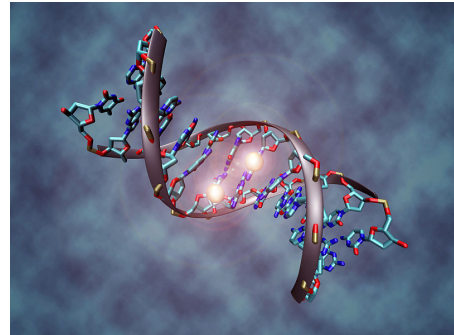
## 1.1. Overview

In Chapter 2, we give a short overview of epigenetic mechanisms of regulation, existing experimental technologies for measuring epigenetic modifications and bioinformatic tools that are commonly used for epigenetic data analysis. In Chapter 3, we introduce EpiExplorer, a tool for intuitive interaction with and visualization of large genetic and epigenetic datasets. We discuss the methodology, as well as the implementation details. We present three use cases: first, we use EpiExplorer to rediscover known properties of CpG islands (see Section 3.4.1). Then, we explore a novel epigenetic mark, *5-hydroxymethylation* (5hmC), together with reference genetic and epigenetic maps and we identify a subset of 5hmC hotspots suitable for experimental validation (see Section 3.4.2). Third, we report on the use of EpiExplorer with cohorts of patient data representing DNA breakpoints from seven cancer types and identify epigenetic and genetic properties of recurring breakpoints (see Section 3.4.3). Finally, we present user statistics that are informative of the impact of our tool in the bioinformatic community (see Section 3.4.5).

In Chapter 4, we present the design and implementation of the EpiGRAPH backend that addresses the computational needs of the EpiGRAPH service. Chronologically, EpiGRAPH was designed and implemented before EpiExplorer, but in this thesis, from a data workflow point of view, I present it after EpiExplorer. In Section 4.3.1, we demonstrate the benefits of using EpiGRAPH and EpiExplorer together by statistically validating hypotheses inspired by EpiExplorer on the association between DNA methylation and sequence patterns. Next, we look into how the association between DNA sequence and DNA methylation varies in different tissues (see Section 4.3.2). Last, we use EpiGRAPH to look at the evolution of DNA sequence and DNA methylation in gene promoters orthologous between mouse and human (Section 4.3.3).

Finally, in Chapter 5, we summarize the work, comment on its impact in the field and provide an outlook into the future challenges and possible developments of our tools.

## 2. Computational epigenetics and text retrieval



Visual representation of DNA methylation by Christoph Bock

### 2.1. The genome. Genomic annotations.

Every mammalian organism starts from a single cell, the *zygote*. The zygote comprises the genome of the organism in the form of several strings of deoxyribonucleic acid (*DNA*). The DNA is organized into a double-stranded helix structure, each strand consisting of a long sequence of *nucleotides*. There are four nucleotides: *adenine*, *cytosine*, *guanine* and *thymine*, noted using the letters *A*, *C*, *G* and *T*, respectively. The two strands of the genome run in parallel and every nucleotide on one of the strand is paired with a nucleotide on the other: guanine is always matched with cytosine and thymine is always matched with adenine.

To ensure optimal packing as well as control accessibility, the DNA is organized into a multi-layered structure. The basic unit of DNA packaging is called *nucleosome* and consists of the DNA wrapped about two-and-a-half times around eight *histone* proteins. After further folding, the nucleosomes are organized into *chromosomes*.

In the course of development of the organism, the cells replicate and differentiate. The DNA is copied (near perfectly) to every new cell. As a consequence, every cell in the mammalian genome has (almost) identical DNA.

Most cellular processes are carried out by biological macromolecules called *proteins*, which consist of more or less long chains of amino acid residues. The amino-acid sequence of each protein is encoded into the DNA, by sequences of nucleotides called *genes*.

DNA and histones together with other proteins involved in cell regulation make up the chromatin. The chromatin is a complex structure, organizing three billion basepairs of

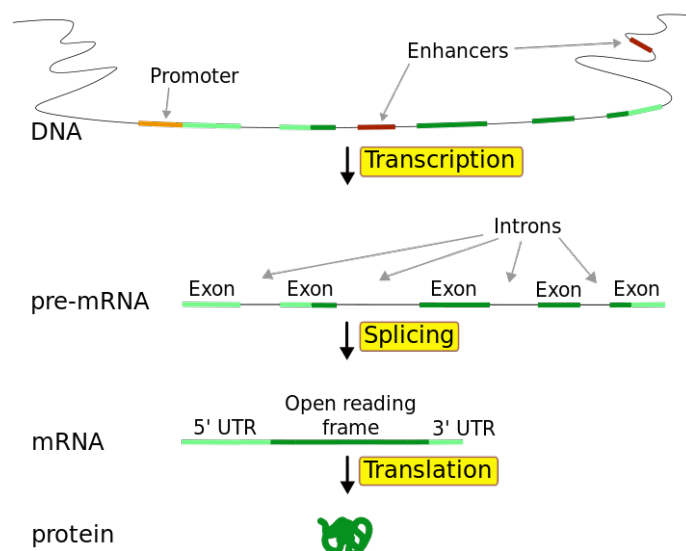


Figure 2.1.: Schematic representation of gene structure and transcription process.  
 From Wikipedia (<http://upload.wikimedia.org/wikipedia/commons/a/a7/Gene2-plain.svg>)

DNA in human, for instance. Its organization and function, even though intensively investigated, is still not completely understood. In their attempts to answer questions related to the functioning of the DNA and related molecular processes, biologists and bioinformaticians typically extract features of the DNA, that are informative of some function. Computational tools commonly encode these features as sets of genomic regions, or *genomic annotations*. Below we list several annotations that we refer to throughout this thesis:

- **Genes, promoters, introns, exons.** Each gene region consists of the following elements: a *promoter*, a *transcription start site*, *exons* and *introns* (see Figure 2.1). Genes are *transcribed* into proteins by the *RNA polymerase*. The RNA polymerase first binds to the gene promoters, separates the DNA strands, and beginning from the transcription start site, produces RNA (ribonucleic acid) complementary to the DNA strand. The resulting RNA undergoes splicing. During splicing a select subset of the exon sequences comprising the splice variant forms a messenger RNA that can be used by the cell to synthesize a protein. The exon sequence used to synthesize the protein is referred to as *coding sequence*.
- **Repeats.** A large part of the mammalian genomes comprises *repeat elements*, or *repeats*. These are patterns of DNA sequence that occur multiple times across the genome. DNA repeats fall into two main classes: *tandem repeats*, where multiple copies of a sequence appear next to each other and *interspersed repeats*, where multiple copies of the same sequence appears in different locations in the genome.
- **Enhancers, silencers, insulators, transcription factor binding sites.** Transcription levels and accessibility of each gene can be influenced by multiple noncoding

functional elements such as *enhancers*, *insulators* and *transcription factors* (TF). Enhancer elements are short regions of DNA, that can attract activator proteins that bind to them and, in turn, recruit mediator complexes. A mediator complex attracts RNA polymerase II which facilitates transcription. Another type of gene regulatory regions are *silencers*. They are the opposite of the enhancers as they inhibit the transcription of nearby genes. Just as enhancers, silencers do not influence the gene directly but attract proteins (*repressors*) that directly or via additional protein complexes interfere with gene transcription. Often when two nearby genes have different transcription patterns, it is important the enhancing or silencing mechanisms that influencing one should not interfere with the other. *Insulator* sequences binded by the CTCF protein serve as genomic barriers that block interactions across them. An important type of regulatory mechanisms are the transcription factors and their DNA binding sites. Transcription factors are proteins that bind to specific DNA sequence patterns found on the DNA (TFBS). Bound transcription factors in turn can interact with additional proteins and implement various regulatory functions.

- **CpG islands.** A cytosine followed by a guanine in the DNA sequence is called a *CpG dinucleotide*. The CpG dinucleotide pattern is underrepresented in mammalian sequence, but tends to cluster in short CpG-rich regions called *CpG islands* (in this manuscript, we also refer to them as CGIs) (Deaton and Bird, 2011). CpG islands are characterized by high guanine and cytosine content (G+C) and elevated CpG occurrences. These criteria, together with a condition on the minimum length of the sequence are how CpG islands are traditionally defined (Gardiner-Garden and Frommer, 1987; Takai and Jones, 2002). CpG islands are important functional regions of the genome, being often a target location for transcription regulation via DNA methylation. DNA methylation involves a methyl group that attaches to an individual cytosine nucleotide usually in the CpG context. An important cause of the underrepresentation of the CpG pattern is *CpG decay* as a consequence of *DNA methylation* (Feuerbach, 2014). Specifically, methylated cytosines (C), due to their biochemical properties, are prone to deaminate into thymine (T) (Holliday and Grigg, 1993). Unmethylated cytosines, on the other hand, deaminate into the nucleotide uracil (U) which is easily recognized and repaired by the DNA repair mechanisms. This is not the case when a methylated C deaminates into a T, leaving DNA repair mechanisms to repair a T/G mismatch. As a result, the T/G mismatch is either repaired to an unmethylated C/G pattern, or wrongly ‘repaired’ into a A/T, thus stably mutating the cytosine into a thymine. Thus, in methylated CpG islands, the methylated CpG dinucleotides are often lost leading to the slow decay of the islands. CpG islands are often found in gene promoters. The reverse association is also observed, as more than two-thirds of all gene promoters co-localize with a CpG island (Deaton and Bird, 2011).

All genomic functional elements discussed above are DNA sequence-based and are identical in all cells. Naturally, complex organisms need to employ dynamic regulatory mechanisms as the cells in different tissues express different proteins and thus need to read off different parts of the genome. A family of regulatory mechanisms that afford such dynamic regulation without changing the DNA sequence are *epigenetic* regulatory mecha-

nisms, which are the main subject of biological research in this thesis.

## 2.2. Epigenetics

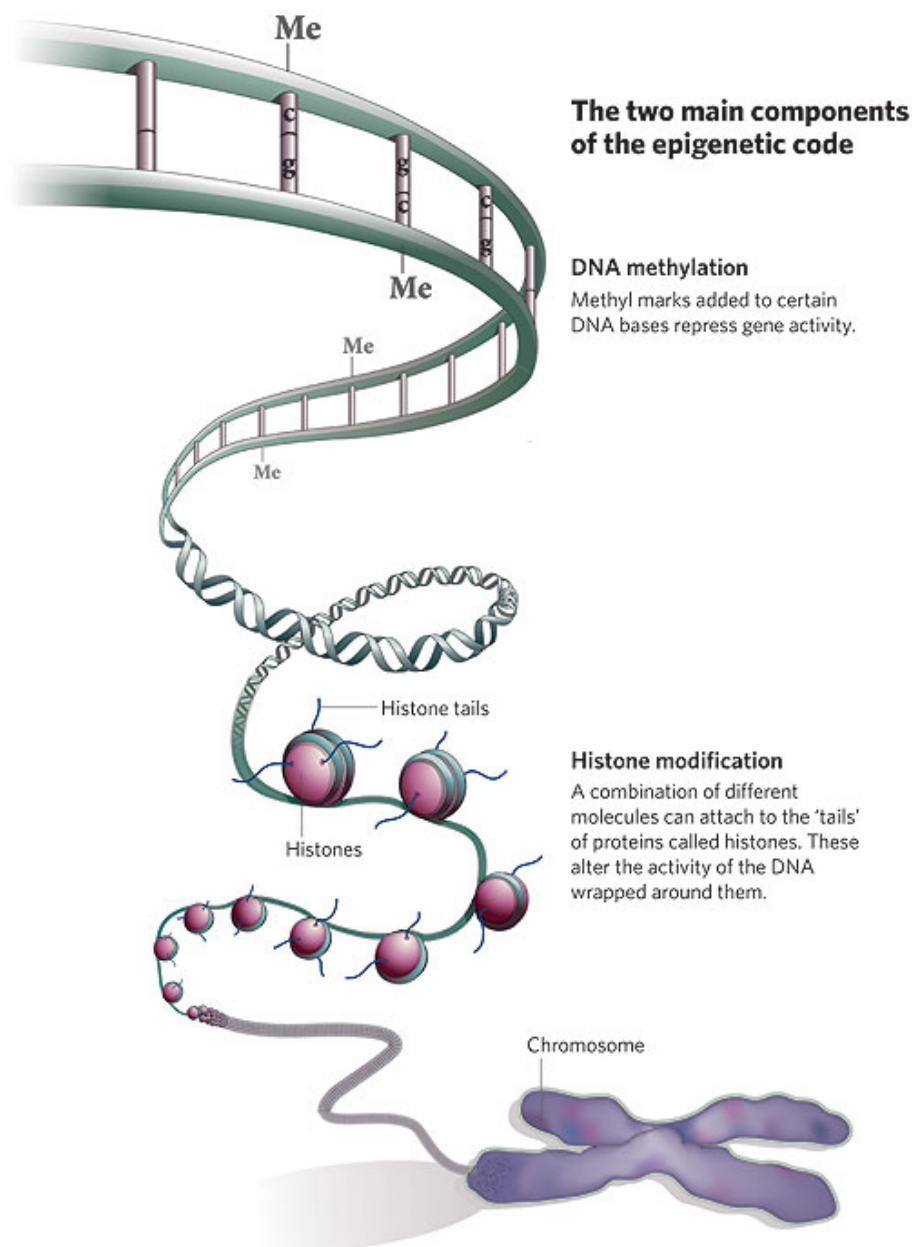


Figure 2.2.: Representation of DNA methylation and histone modifications, the two main actors in epigenetic regulation. Reprinted from Qiu (2006))

In Greek, 'epigenetics' translates as 'on top of genetics'. The term was introduced in the early 1940s by Conrad Waddington (Waddington, 1942). More than 70 years later, there are many definitions of epigenetics, but there is still no established one. A commonly used definition is that epigenetics is 'the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence' (Russo

et al., 1996). A more recent proposal is that ‘epigenetic state is the structural adaptation of chromosomal regions so as to register, signal or perpetuate altered activity states’ (Bird, 2007). Berger et al. (2009) propose the following definition: ‘an epigenetic trait is a stably heritable phenotype resulting from changes in a chromosome without alterations in the DNA sequence’. In a broader sense, epigenetics connects the genome and the phenotype by empowering the same genetic information to give rise to cells with different identities within the same organisms. There are several well known epigenetic mechanisms, among which: *DNA methylation*, *chromatin modifications*, *RNA-induced silencing*. In this thesis we focus on the first two, as they are to date the most studied epigenetic mechanisms (see Figure 2.2). In the following sections, we present them in detail.

## 2.3. DNA methylation

DNA methylation is the only epigenetic modification that targets directly the DNA sequence. In vertebrates, DNA methylation refers to a methyl group attaching (mostly) to cytosines in the context of CpG dinucleotides. In fact, 60%-80% of all CpG dinucleotides in the human genome are methylated (Smith and Meissner, 2013). Unlike CpG dinucleotides, the majority of CpG islands tend to be consistently *unmethylated*. The methylation of promoter CGIs has been linked to the transcriptional activity of the associated genes. Methylated CpG islands in gene promoters have been demonstrated to be associated with transcriptional silencing, while the unmethylated state is associated with transcriptional activation (Deaton and Bird, 2011). Notably, for gene promoters that do not overlap with CpG islands, the DNA methylation of the present CpG dinucleotides does not seem to influence the gene expression. These observations hint at the important regulatory role of DNA methylation.

### 2.3.1. Functions and mechanisms

DNA methylation plays an important role in regulating gene transcription, early embryonic development and cell differentiation, a number of diseases and aging. Specifically, DNA methylation increases or decreases the binding affinity of specific proteins to the DNA sequence in question (Jaenisch and Bird, 2003). A well studied example is suppressing imprinted gene copies and DNA sequences with latent transcriptional activities, such as transposable elements (Smith and Meissner, 2013).

DNA methylation guides epigenetic regulation during early embryonic development and dynamically changes in the course of cell differentiation (Reik, 2007). A number of gene promoters have been identified to possess tissue-specific methylation pattern associated with their expression patterns (Song et al., 2005; Illingworth et al., 2008). That is, genes have been shown to be unmethylated in some tissue and methylated in another. Furthermore, housekeeping genes are constitutively unmethylated in all tissues. To demonstrate the causal relationship between methylation and transcription, gene transfer experiments were performed in which unmethylated genes were actively transcribed and if the same genes were remethylated transcription was inhibited (Cedar and Bergman, 2012).

DNA methylation plays an important role in suppressing transposons. Transposable elements, (also called transposons, or jumping genes) are DNA sequence segments that have

the ability to move across the genome. They often duplicate and as a consequence are a major type of DNA repeats, which constitute nearly 40% of the mammalian genomes. The three main types of DNA repeats associated with transposable elements are long interspersed nuclear elements (*LINEs*), short interspersed nuclear elements (*SINEs*) and long terminal repeats (*LTR*). Transposons often contain a promoter that needs to be repressed in order to prevent their transcriptional activity that can be lethal to the cell. In adult cells, these regions and especially their promoters are consistently hypermethylated (Smith and Meissner, 2013).

DNA methylation has long been associated with establishing and maintaining of genomic imprinting (Li et al., 1993). Imprinting is a process by which several dozens genes are transcribed only from the paternal (or maternal) genome and silenced in the other (Reik and Walter, 2001). After they are formed, the imprinted methylation marks are preserved during the developmental cycle and are faithfully transferred during each mitosis.

DNA methylation has also been indicated as an important factor in diseases, including in the progression of multiple cancer types (Esteller, 2007). In addition to the coordinated changes during normal development, in cancer the DNA methylome exhibits characteristic changes. Such abnormal methylation patterns have been used to diagnose functionally compromised cell states (Esteller, 2007). These may include both genome-wide hypomethylation or hypermethylation or more specific targeted aberrant gene promoter methylation. In particular, hypermethylation of promoters of tumor-suppressor genes can lead to their transcriptional suppression and result in abnormal cell functions (Bernstein et al., 2007).

Mechanistically, methylated cytosines promote or prevent the access to a genome loci of regulatory proteins (Bernstein et al., 2007). Methylated cytosines can mediate transcriptional repression if bound by a family of methyl-binding proteins. These proteins removes acetylation from the chromatin-forming histone proteins thus reducing chromatin accessibility (Bird, 2002). DNA methylation has been shown to influence the chromatin structure and thus the accessibility of the DNA sequence. A particular convincing evidence is when an unmethylated DNA sequence after its integration in the genome is packed into an open chromatin structure, while if the same sequence is artificially methylated, it is packed into a closed chromatin (Keshet et al., 1986). Alternatively, DNA methylation of a transcription factor binding site can prevent binding, as observed for the CTCF protein (Bernstein et al., 2007).

### **2.3.2. The dynamics of DNA methylation during embryonic development and tissue differentiation**

DNA methylation undergoes two distinct epigenetic *reprogramming phases* (see Figure 2.3). The first phases occurs early during embryonic development, as DNA methylation is actively erased genome-wide (Reik et al., 2001). This is followed by a wave of de novo methylation during which genome-wide methylation is reestablished (Reik, 2007). During this phase, pluripotency-associated genes, such as Oct4 and Nanog are reactivated. As the cell differentiates, pluripotency-associated genes are silenced in order to protect the differentiated cells from undesired return to pluripotency and tissue-specific genes are activated (Epsztejn-Litman et al., 2008). Once established, DNA methylation patterns in differentiated cells are believed to be generally stable (Reik, 2007). Not all DNA methylation marks



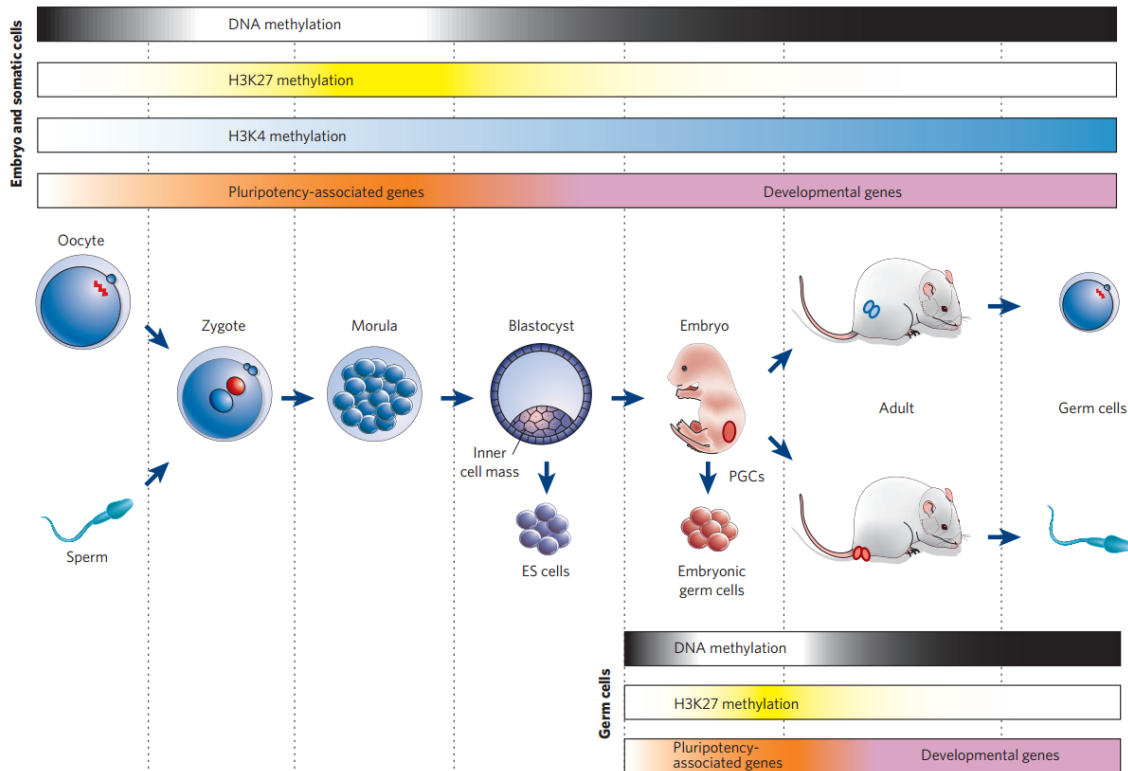


Figure 2.3.: Schematic representation of the two reprogramming stages of DNA methylation and histone modifications. Figure reprinted from Reik (2007)

are erased during the early genome-wide DNA demethylation. For example, CpG islands associated with transposons need to be stably silenced in order to protect the host genome. Also, the repressive marks of imprinted genes are protected during the developmental reprogramming and are preserved for all differentiated cells. These marks are only erased during the second reprogramming phase in primordial germ line cells, the cells that give rise to the germ line. The second phase of epigenetic reprogramming resets DNA methylation and re-establishes pluripotency in germ line cells (Laurent et al., 2010). An important function of that phase is to remove the methylation marks of imprinted genes, which are later re-established in a parent-specific manner (Cedar and Bergman, 2012). Outside these two major reprogramming events, the development and differentiation is characterized by stable somatic inheritance of DNA methylation, low mutation rate, differential methylation of tissue-specific genes and stable silencing of pluripotency-associated genes, imprinted genes and transposon elements.

The dynamics of DNA methylome discussed previously are only possible with efficient DNA methylation deposition and maintenance. Three DNA methyltransferase enzymes are responsible for this: Dnmt1, Dnmt3a, Dnmt3b (Li et al., 1992; Okano et al., 1999; Jones and Liang, 2009). Dnmt1 is a maintenance methyltransferase, as it propagates DNA methylation during replication (see Figure 2.4). Dnmt3a and Dnmt3b are responsible for de novo DNA methylation (Smith and Meissner, 2013) and play a critical role in re-establishing of DNA methylation in early embryonic development (see Figure 2.3). A number of experiments were performed to better understand the dependencies between

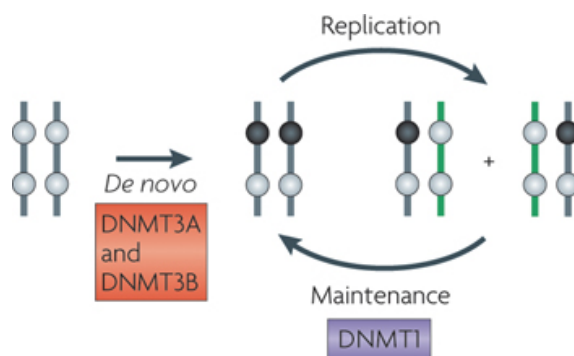


Figure 2.4.: DNA methylation deposition and maintenance by DNA methyltransferases Dnmt1, Dnmt3a and Dnmt3b. Figure reprinted from Jones and Liang (2009)

DNA methylation, pluripotency and differentiation of embryonic stem cells (ESC) (Smith and Meissner, 2013). For example, ESC deprived of Dnmt1 experience a rapid loss of DNA methylation until they stabilize with around a fifth of the DNA methylation of normal ESC. The cells are stable, but die of apoptosis when induced to differentiate. If Dnmt1 is reintroduced in those cells then DNA methylation is restored with the exception of imprinted patterns (Tucker et al., 1996). If Dnmt3a and Dnmt3b are knocked out in ESC, the cells lose almost all DNA methylation after multiple divisions (Chen et al., 2003; Jackson et al., 2004).

DNA methylation patterns are primarily maintained through cell division via Dnmt1. The so-called maintenance methyltransferase targets hemi-methylated CpG dinucleotides and methylates the unmethylated strand. In that way, it efficiently preserves DNA methylation during cell division as it propagates any DNA methylation to the daughter cells (Bernstein et al., 2007). Dnmt1 is constantly found at the replication foci (Leonhardt et al., 1992) thus ensuring the consistent replication of DNA methylation pattern across the genome.

Unlike mechanisms for de novo methylation and maintenance, mechanisms for DNA demethylation are less understood. Demethylation occurs globally in two waves during early development and gametogenesis. Alternatively, site-specific demethylation occurs, often involving tissue-specific regulation. DNA demethylation is hypothesized to work in three main modes: active demethylation, passive demethylation by cell division and demethylation by repair (Cedar and Bergman, 2012). The exact specifics of DNA demethylation are not entirely clear, but one of the proposed workflows suggests an initial conversion of methylated cytosine (5mC) to 5-hydroxymethylated cytosine (5hmC), which in turn is converted to 5-hydroxymethyluridine (5hmU) and subsequently converted to a cytosine by DNA repair (Cedar and Bergman, 2012). Recent studies analyze the specifics of how 5mC can be converted to 5hmC that in turn results in demethylation, a process often catalyzed by a ten-eleven translocation (TET) family of proteins (Ito et al., 2010; Wu and Zhang, 2010). An interesting experiment showed that reducing the activity of the Tet1 protein results in increased DNA methylation in CpG islands (Wu et al., 2011; Ficz et al., 2011). These results have led to the hypothesis that the genome-wide DNA methylation is even more dynamic than previously believed and involves continuous resetting through 5hmC

(Deaton and Bird, 2011).

Active demethylation may also involve a *deamination* step (Wu and Zhang, 2010). A link between deamination and DNA demethylation has been shown experimentally, as Popp et al. (2010) report increased DNA methylation levels in embryos without activation-induced deaminase (AID) enzyme. AID stimulates the deamination of methylated cytosines in a CpG pattern to uracils that often are wrongly repaired to thymine, resulting into a TpG pattern. As we discuss later in this thesis, in methylated CpG islands the TpG dinucleotide has a higher frequency (and its reverse complementary CpA), which are the dinucleotides methylated CpGs deaminate into.

### 2.3.3. Experimental technologies

Below we present an overview of the experimental technologies available for measuring DNA methylation.

Available sequencing and microarray technologies cannot recognize DNA methylation directly because standard molecular techniques such as polymerase chain reaction (PCR) erase the DNA methylation marks (Laird, 2010). As a consequence, methods were developed that translate DNA methylation information into DNA sequence information based on the chemical properties. These are often combined with high-throughput sequencing or microarray technologies (Bock et al., 2010a; Bock and Lengauer, 2008). In this section, we briefly discuss the *bisulfite sequencing*, *microarrays* and *enrichment-based methods* (Bock, 2012).

The gold standard for determining if a cytosine is methylated is bisulfite sequencing (see Figure 2.5). The method uses the following property: when DNA is treated with sodium bisulfite, the unmethylated cytosines are converted to uracil (U), while the methylated cytosines remain. These are processed by a standard genetic sequencing procedure, yielding either thymine (unmethylated C) or cytosine (methylated C). In short, the method converts the methylation mark into a sequence difference that can be identified by comparison with a reference sequence. The advantage of bisulfite sequencing is that it provides exact measurements of methylation for every CpG. The quality of the data comes at a cost, as bisulfite methods struggle to scale to the genome (Eckhardt et al., 2006). To address this shortcoming, bisulfite sequencing has been combined with enriching of specific enzymes (as in reduced representation bisulfite sequencing - RRBS (Meissner et al., 2008)) allowing to target large parts of the genome at significantly reduced cost (Gu et al., 2010). While cost-effective and providing genome coverage, the specificity of the enzyme targeting do not afford truly genome-wide analysis of DNA methylation patterns due to the biases of the data sampling, such as sampling predominantly from CpG-rich regions. As of recently, it was also discovered that bisulfite sequencing struggles to distinguish between methylated and hydroxymethylated cytosines (Huang et al., 2010; Bock, 2012).

An alternative to the previous method combines bisulfite treatment of DNA methylation with custom-designed microarrays. Similarly to the RRBS, the costs to obtain genome-scale measurements are affordable. However, the disadvantages include the need for a complicated initial array setup, and again access to only limited part of the genome (Bibikova et al., 2009). Illumina offers popular custom arrays to interrogate DNA methylation in human. The Illumina GoldenGate BeadArray covers around 1,500 CpG sites, the Illumina

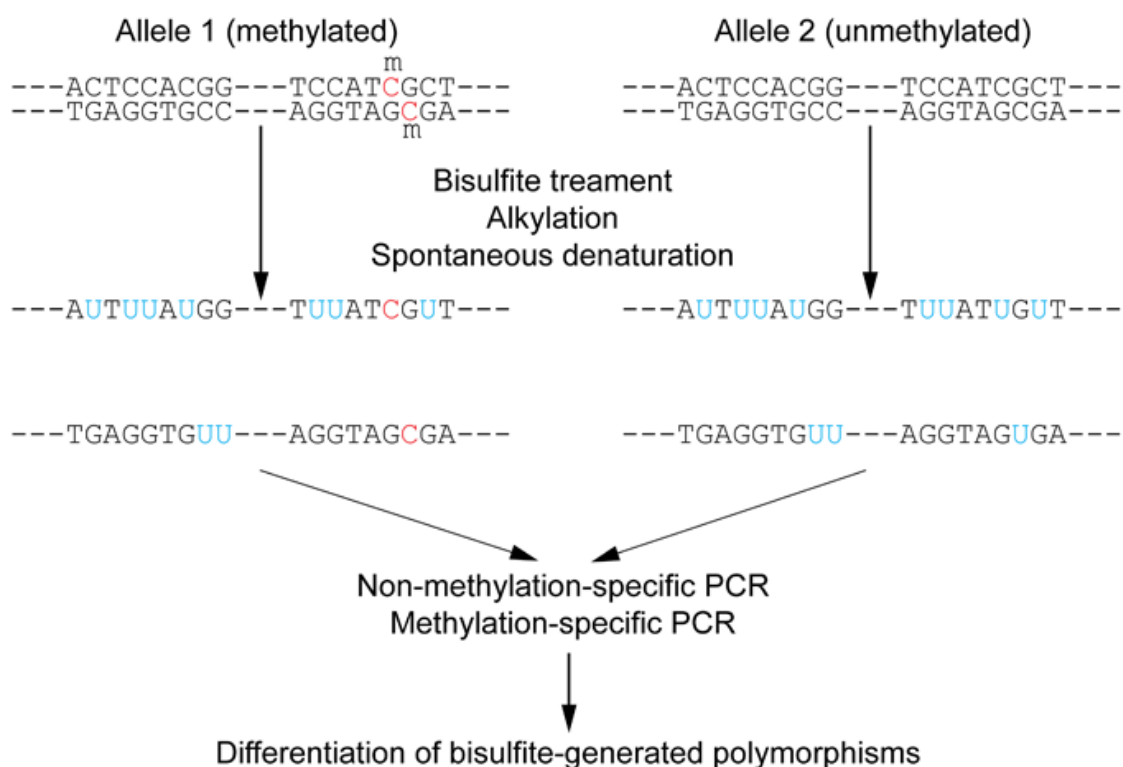


Figure 2.5.: Schematic representation of bisulfite sequencing from Wikipedia ([http://upload.wikimedia.org/wikipedia/en/c/c9/Wiki\\_Bisulfite\\_sequencing\\_Figure\\_1\\_small.png](http://upload.wikimedia.org/wikipedia/en/c/c9/Wiki_Bisulfite_sequencing_Figure_1_small.png))

Infinium assay covers more than 27 thousand CpG sites (Laird, 2010) and the Illumina Infinium HumanMethylation450 that offers interrogation of more than 485,000 methylation sites.

An alternative to bisulfite sequencing for measuring DNA methylation are the enrichment-based strategies. These methods identify DNA regions with methylated or unmethylated DNA and use next-generation sequencing to extract the sequences (Bock et al., 2010a). When using MeDIP-seq (Weber et al., 2005), antibodies are used that bind to methylated DNA fragments. With MethylCap-seq (Brinkman et al., 2010), a methyl-binding domain is used to identify domains with similar levels of DNA methylation. Alternatively, restriction enzymes can be used to identify only DNA sequences with high levels of DNA methylation (Brunner et al., 2009). Generally, enrichment-based strategies succeed where the bisulfite sequencing methods fail and vice versa. Namely, one can perform genome-scale experiments at a relatively low cost, however this comes at the cost of the lack of information about the CpG-poor regions combined with lower quality of the experimentally measured levels of DNA methylation. A final advantage of these methods is the ability to distinguish between 5mC and 5hmC (Bock, 2012).

## 2.4. Histone modifications

DNA is wrapped around nucleosome molecules. A nucleosome is an octamer formed of four pairs of histone proteins: H2A, H2B, H3 and H4. Each histone protein has a long unstructured tail at the end of its amino-acid chain. The amino acids on the histone tails are often subject to various chemical modifications. These include methylation, phosphorylation and ubiquitination. Unlike DNA methylation, histone modifications regulate the access to the DNA by influencing the compaction level of the chromatin. Specifically, they do so not only by their presence and affecting inter-nucleosomal interactions, but also by recruiting of remodelling enzymes and other proteins and complexes with specific enzymatic activities (Bannister and Kouzarides, 2011). Open *euchromatin* increases the accessibility of the DNA for binding, while tightly packed *heterochromatin* has the opposite effect. Heterochromatin covers large sections of the genome and is considered to have evolved as a highly repressive structure that limits the access to the underlying DNA sequence, for example to prevent the activation of transposable elements (Beisel and Paro, 2011). Most importantly, histone modifications (and chromatin-based silencing) is propagated through DNA replication, for example through the activity of the Polycomb-group (PcG) and trithorax-group (trxG) protein complexes, ensuring the consistency of epigenetic regulation between cells (Beisel and Paro, 2011).

### 2.4.1. Dynamic regulation via histone modifications

Histone modifications regulate transcription via two main mechanisms: either by directly affecting the chromatin structure or by managing (recruiting or preventing) the access of specific proteins to the DNA sequence (Bannister and Kouzarides, 2011). The latter are better characterized as various proteins have been identified that interact with specific sequence domains depending on the presence or absence of particular histone modifications (Xhemalce et al., 2011). For the former, one of the better understood examples is how histone acetylation by reducing the electrostatic interactions between the DNA molecules and the histone proteins leads to open chromatin (Bannister and Kouzarides, 2011).

There are more than 100 post-translational histone modifications. They are annotated based on the histone protein on which they occur, the amino acid they influence and its position on the histone tail and the type of the modification. These are commonly annotated with short abbreviations. For example, H3K9ac would indicate an acetylation of the H3 histone at the lysine at the 9th position. Similarly, H3K27me3 refers to a lysine on the 27th position of the H3 with observed trimethylation. A specific lysine can have up to three methylation group attached and the resulting function can differ depending on the number of methyl groups. For example, monomethylation of H3K4 is associated with marking of enhancer elements, while trimethylation on the same position is generally associated with active unmethylated CpG island promoters. Furthermore, trimethylation at different positions can have different regulatory function. For example, H3K27me3 is often associated with repression, while H3K4me3 is associated with transcriptional activity.

Trimethylation of H3K4 is often observed at unmethylated CpG-island gene promoters (Bernstein et al., 2007). The association between H3K4me3 and unmethylated CpG islands may be the result of the activity of trithorax-group (trxG) protein complexes, associated

with maintaining gene expression, that bind to unmethylated DNA and catalyze H3K4 methylation (Bernstein et al., 2007). However, several experiments have shown that the modification does not influence gene transcription directly. For example, the depletion of the H3K4 methyltransferase complex results in globally reduced H3K4me3, but does not result in immediate difference in transcriptional activity (Zentner and Henikoff, 2013).

Monomethylation of H3K4 has been associated with putative enhancers (Heintzman et al., 2009; Zentner et al., 2011). Additionally, enhancer sequences often co-localize with trimethylation or acetylation marks on H3K27 (Zentner and Henikoff, 2013). As we will show later in this thesis, H3K4me1 often co-localizes also with DNA 5-hydroxymethylation (5hmC), often in the context of enhancers (Section 3.4.2)

The trimethylation of H3K27 has been a subject of a lot of research and is classically associated with closed chromatin. H3K27me3 is often observed in repeat regions potentially repressing their activity. However, in ES cells the repressive H3K27me3 modification is often discovered together with the activating mark H3K4me3 in gene promoters. As the ES cells differentiate, these so called bivalent promoters often lose one of those marks and remain repressed or activated in adult cells. H3K27me3 displays a variety of functions in dynamic tissue-specific transcriptional regulation (Bernstein et al., 2007).

Unlike the variance in the regulatory functions of histone methylation, lysine acetylation generally is associated with euchromatin and transcriptional activity. The exact mechanism involves reducing the positive charge of the histones, leading to less tight interaction between the DNA and the histones preventing closed chromatin and in effect leading to open chromatin (Kouzarides, 2007). This has also been observed for histone phosphorylation (Bannister and Kouzarides, 2011). The challenges of understanding the dynamics of histone regulation increase as different histone modifications can influence each other (Bernstein et al., 2007)

Most histone modifications are directly applied by specific enzymes. Many of these enzymes are identified, such as for methylation (Zhang and Reinberg, 2001) and for acetylation (Stern and Berger, 2000). Moreover, the enzymes that remove the modifications are also identified (Kouzarides, 2007). Histone modifications also employ complex mechanisms to ensure their epigenetic inheritance. These often involve recruiting multiple proteins to ensure the preservation of the histone marks on the daughter cells. For example, the heredity of H3K4 and H3K27 methylation are mediated by Polycomb-group (PcG) and trithorax-group (trxG) protein complexes (Bernstein et al., 2007). The H3 and H4 histones are distributed randomly between the new cells, thus ensuring that the DNA of both cells is packed in chromatin with the same histone modifications, further propagated by PcG and trxG (Zentner and Henikoff, 2013).

## 2.4.2. Experimental technologies

Chromatin immunoprecipitation (ChIP) remains the main method to identify histone modifications (Figure 2.6). The standard ChIP protocol involves shearing the chromatin (by sonication) to generate only DNA segments of small size and then binding antibodies against a specific histone modification to isolate DNA segments by precipitation (Bernstein et al., 2007). Afterwards, the precipitated DNA sequence fragments can be analyzed using different methods. Similarly to bisulfite sequencing, ChIP can be combined with microar-

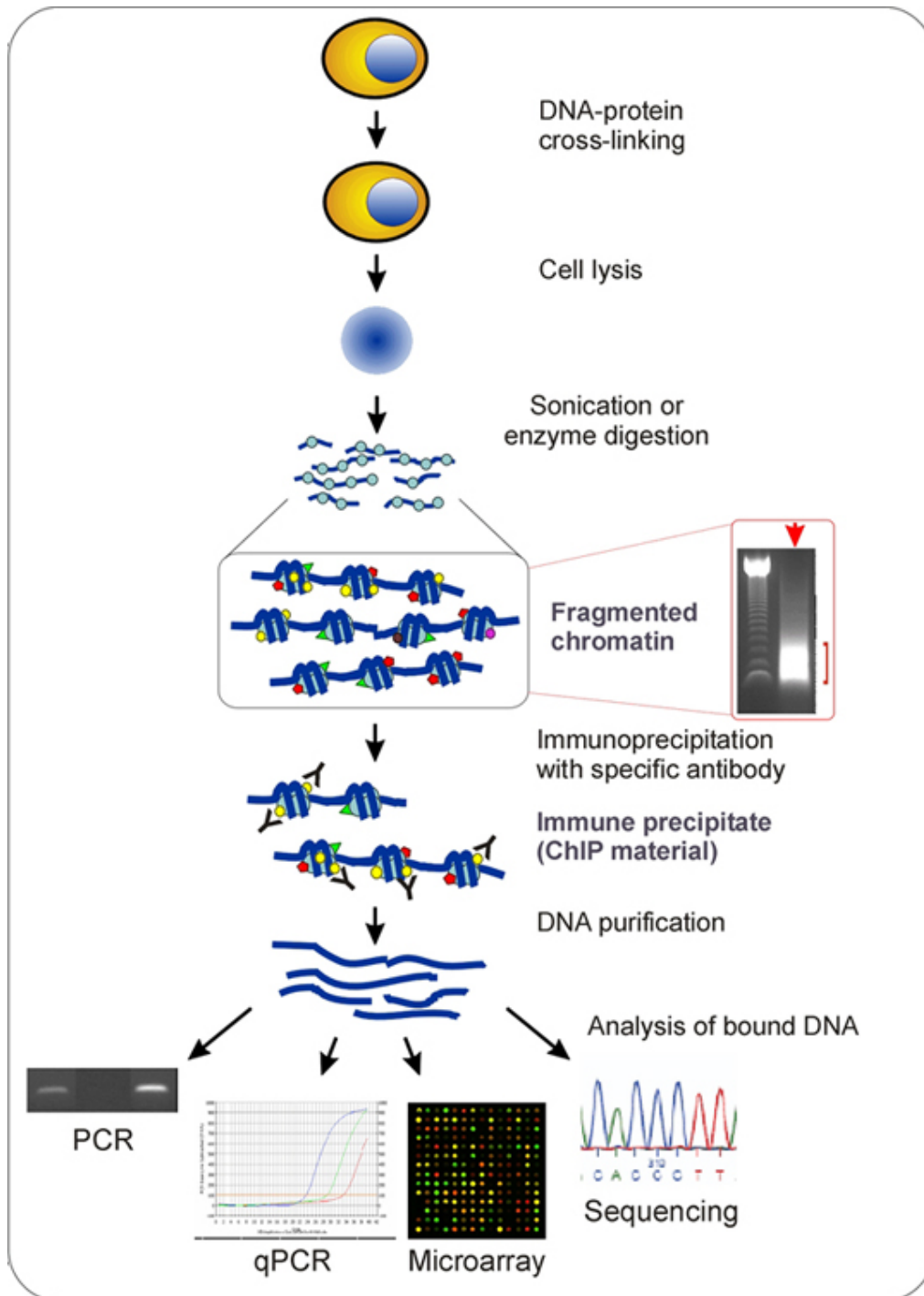


Figure 2.6.: Schematic representation of chromatin immunoprecipitation. Figure adapted from Collas and Dahl (2008)

rays (ChIP-chip) and with high throughput sequencing (ChIP-seq) to provide genome-wide maps. The main issues of the method are related to the quality of the used antibody. Problems such as the specificity of the antibody binding and potential disrupting when binding neighboring nucleosomes can lead to underestimation and biases in the estimations of the presence of specific histone modifications (Kouzarides, 2007).

## 2.5. Computational analysis of genomic and epigenomic data

Hawkins et al. (2010) suggest that computational biology needs tools that offer integrated analyses on all the various experimental data types. Goecks et al. (2010) propose bioinformatic tools to support accessible, reproducible and transparent research. In what follows, we give a short overview of the existing types of tools for working with epigenome data and how they meet a variety of criteria.

### 2.5.1. Overview of bioinformatic tools for working with epigenome annotations

- **Tools for data preprocessing and quality control** In the previous sections, we discussed various methods for extracting DNA methylation and histone modification data. After the raw data is obtained, a processing and quality control step is typically performed (Bock, 2012). The main goal of the data processing is to produce comparable (among different experiments) and accurate absolute values for the correct genomic positions. In the case of DNA methylation, the purpose is to extract absolute and comparable DNA methylation levels for every CpG. In the case of histone modifications, the result is a map with observed occurrences for every genomic position. These values are often subject to a peak-identification algorithm to precisely indicate genomic regions where the modification is located. The data processing techniques vary, depending on the technology used. During the quality control step, one keeps an eye on a number of technical details and ensures that potential biases due to the study cohort, sample material, experimental protocol, batch effect or others do not influence the quality of the data. In this thesis, we work directly with already processed and quality-controlled data and do not go into further details regarding the existing methods. However, interested readers can find a detailed overview of the preprocessing and quality control methods in Bock (2012).
- **Integrated databases and genome browsers** Epigenetic data that is mapped onto the genome can be readily visualized in alignment with other established reference genomic and epigenomic maps. Genome browsers – probably the most known and widely used bioinformatic tools – serve this purpose, to navigate the user through data aligned to the genome. ‘Genome browsers’ is a generic name by which the classic UCSC Genome Browser (Karolchik et al., 2011) and Ensembl (Flicek et al., 2010) are known, as well as the new WashU Human Epigenome Browser (Zhou et al., 2011) and Integrative genomics viewer (Robinson et al., 2011). They provide an intuitive visual interface that represents the genome as a one-dimensional map, onto which data from multiple annotations are overlaid. The users can zoom into and out of any genomic region and observe interesting local associations between annotation tracks. The power of the genome browsers lies in their easy-to-use, intuitive interface and the extensive annotation database that they support. They are often used in studies focused on the properties of one or several genomic loci. The main shortcoming of browsers is their inability to provide a simultaneous analysis of a set of regions, and thus to generalize observations outside the scope of a (narrow) region.
- **Genome calculators** An important family of computational tools are designed for performing common operations, such as filtering region sets to comply with cer-



tain properties, intersection between region sets, computing genome coverages etc. UCSC Table Browser (Karolchik et al., 2004) and Ensembl BioMarts (Kinsella et al., 2011) provide user-friendly interfaces for selecting and refining datasets based on different criteria. They essentially provide interfaces to the databases underlying the genome browsers. Command-line tools, such as BEDTools (Quinlan and Hall, 2010) and BEDOPS (Neph et al., 2012) are often used to perform these operations via a command-line interface and, while they lack user-friendly interfaces and intuitive data visualizations, they can be easily integrated into workflows. Finally, scripting and programming languages, such as R/Bioconductor (Gentleman et al., 2004) and Python (Python Programming Language, 2009), allow great flexibility of the operations, but require extensive programming experience. All these tools suffer from the same drawback: it is relatively difficult to discover interesting aspects in new data; they lack quick visualizations and exploratory output; rather, it takes large effort and computer skills to investigate single aspects of the data, which have to be thought of and planned ahead.

- **Workflow tools** Basic operations with genomic annotations, as discussed above, are often an integral part of visual workflow toolkits such as Galaxy (Giardine et al., 2005), its extension the Genomic HyperBrowser (Sandve et al., 2010), Taverna (Hull et al., 2006) and Genome-Space (Reich et al., 2012). They offer intuitive user interfaces that allow complex analysis pipelines, but often require careful planning of each data analysis.
- **Statistical analysis and prediction** There are very few tools that offer a generalized framework for modeling statistical dependencies between genomic and epigenomic annotations. Among the tools with narrower scope we mention: GREAT (McLean et al., 2010) and the Genomic HyperBrowser (Sandve et al., 2010). GREAT offers insights into the biological function of sets of genomic regions based on their nearby genes. The Genomic HyperBrowser offers a visual interface to define and evaluate statistical association between genomic annotations. In this thesis we present EpiGRAPH, our web service that can be used to define general purpose statistical models on epigenomic and genomic data.

### 2.5.2. The characteristics of existing bioinformatic tools

In Table 2.1, we present a summarized view of the characteristics of bioinformatic tools for analysis of epigenetic data. We formulate a set of criteria that in our view are meaningful for most types of epigenetic data analysis and we appreciate how these criteria are met by the various tools. The criteria are:

- **User interface:** can the user define an analysis via a user interface, or does he or she need knowledge of command-line scripting or programming?
- **Visual results:** are the results of the analysis presented in a visually intuitive way, by meaningful graphics for example?
- **Speed of the analysis:** can the user set up, run the analysis and get the results very fast (within a minute)?

Type	Example	User interface	Visual results	Speed of analysis	Analyzes region sets	Integrated database	Statistical inference	Naïve exploration
Genome browsers	UCSC Genome Browser	✓	✓	✓	✗	✓	✗	✓
Database interfaces	Ensembl Biomarts	✓	✗	✗	✓	✓	✗	✗
Command-line computations	BEDTools	✗	✗	✗	✓	✗	✗	✗
Programming	R	✗	✓	✗	✓	✗	✓	✗
Workflow tools	Galaxy	✓	✓	✗	✓	✓	✗	✗
Enrichment analysis	Genomic HyperBrowser	✓	✓	✗	✓	✓	✓	✗
Statistical analysis	EpiGRAPH - presented in this thesis	✓	✓	✗	✓	✓	✓	✓
Interactive exploration	EpiExplorer - presented in this thesis	✓	✓	✓	✓	✓	✗	✓

Table 2.1.: Characteristics of the most popular bioinformatic tools for analysis of epigenome data.

- **Analyzes sets of regions:** can the tool work with sets of regions and thus provide general views on genomic and epigenomic function, or is it focused on the properties of single regions?
- **Integrate database:** does the tool support or connect to a database with substantial number of reference annotations or the user needs to provide all data?
- **Statistical inference:** can the tool perform statistical testing or statistical modeling in order to rigorously evaluate significance of associations?
- **Naïve exploration** can the tool help the user understand basic characteristics of new, unfamiliar datasets or does it expect prior knowledge on features, name spaces, ranges of values etc?

Note, that the lack of a certain characteristic does not necessary mean that we believe that the tool is inferior in some way. For example, BEDTools may only fit one of our criteria, but is very good at performing basic operations with region sets.

In the upper part of Table 2.1, we observe that there is a lack of tools that offer statistical evaluation of analysis results and real-time interaction. Also, genome browsers are the only tools that do not expect the user to be familiar with the data in a way that she can code the analysis by herself. For example, databases and command-line tools require the user to at least know the tables, value types and ranges of the features. Programming tools can provide information about the data, but the user needs to have advanced computer science skills. In our view, only genome browsers are simple and intuitive enough to help the user familiarize themselves with a new dataset without prior knowledge. However, they lack flexibility to work with sets of regions.

Motivated by those observations, in this thesis we describe two software tools and interactive web servers: EpiExplorer and EpiGRAPH. With EpiExplorer, we developed a method and software that combines the interactive nature of genome browsers with the region-based analytical approach of Galaxy, enabling users to explore large-scale genomic datasets in search of interesting functional associations. EpiGRAPH provides a statistical framework that automatically tests a large number of possible associations and reports to the user the significant findings. The characteristics of the toolkits are represented in the lower part of Table 2.1 in blue.

## 2.6. Text retrieval and exploratory search

### 2.6.1. Information retrieval and text retrieval

The field of *Information retrieval* (IR) is concerned with ‘finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)’ (Manning et al., 2008). *Text retrieval* is a branch of IR that specializes in finding textual information in collections of text documents. Popular examples of text retrieval, which daily engage hundreds of millions of people, are web search engines. Web search engines enable Internet users to find relevant information in billions of web pages. Text retrieval systems expect a short textual query and automatically compute a list of documents ordered by estimated relevance to the query. An IR system aims to efficiently retrieve all documents related to the query and to rank them in decreasing number of relevance. If the ranking of the documents is not of importance, then we refer to the problem as *boolean retrieval*, that is a document either matches the query or not.

An common scenario in information retrieval is *Exploratory Search* (Marchionini, 2006). Exploratory search refers to performing a search when the user is unfamiliar with the domain or unsure about the expected results or the queries needed to achieve them. For example, exploratory search is popular in e-commerce websites, where clients often are unsure of the exact product they are searching for or of the terminology to describe it. Exploratory search is also a common scenario in web search.

Most commonly, web users query search engines to identify the most relevant information about a concept. Web users often have little or no prior knowledge of the topic and use a series of evolving queries to familiarize themselves with it. Similarly to web users, biologists are often interested in finding the relevant genomic and epigenomic properties of a set of genomic regions (for example, maps of enrichment of a novel epigenetic mark) about which they do not have extensive prior knowledge. In web search, instant response is crucial as it enables the user to update and refine a request quickly, thus increasing the knowledge about the topic of interest. In bioinformatics, similar speed of simple queries can be as important when offering a flexible and intuitive interface for probing into epigenetic properties. Fast responses also facilitate quick feedback and easy correction of imprecise queries offering a fault-tolerant service, which is of value especially when dealing with investigating novel biological properties about which little is known.

### 2.6.2. Data indexing and query types

The efficiency of an information retrieval system depends on how its data collection is preprocessed and stored and the types of queries it should answer. Data can be indexed into relational databases, non-relational database or search index structures.

*Relational databases*, such as MySQL (MySQL Database, 1995) and Oracle (Oracle Database, 2009), require the data to be well structured, with associated data types and described by a database schema. The efficiency of a relational database solution depends on optimized design of a database schema to match the data and the queries. Traditional relational databases focus on answering boolean queries (identify all records that match a query) and often do not provide ranking of the matching results. The query language, SQL, allows very flexible and complex queries, but often requires advanced database knowledge. The requirement for a fixed database schema together with complex query language limits the applicability of relational databases to bioinformatic software.

*Non-relational databases* (or *NoSQL* databases) are characterized as semi-structured databases that do not have a schema and supports records with different fields. NoSQL databases fall into several different types: document stores (e.g. MongoDB, MongoDB (2009)), key-value stores (e.g. Redis, Redis (2009)) and wide column stores (e.g. HBase, HBase (2008)). Each type is optimized for different data formats and queries, thus enabling a large range of applications not suitable for relational databases. However, they often do not support JOIN-line queries.

*Search-engine indexes* facilitate full text search with advanced ranking methods. The data structures are optimized for textual data with little or no structure and for easy forming and answering of queries. They often support prefix search (find on-the-fly words that start with a query term) and wildcard search (queries can have wildcard characters allowing flexible text search). Prefix search and wildcard search enable *auto-completion* and *faceted search*

*Auto-completion* – also known as *suggest search* – has been introduced in various services on the web over the last few years, most commonly in web search. It typically consists of an input interface through which the user starts introducing a query. As the input query is being typed, the search engine automatically suggests the most probable completions of the current partial query. The idea has revolutionized the field of text search as the search engine assists the user to form a relevant request. Auto-completion was added to the search engines relatively late (around 2006), mainly because the index structure commonly used for search engines, the inverted index, did not naturally support efficient auto-completion.

*Faceted search* is a technique for accessing and visualizing structured or semi-structured information that can be described by multiple properties. Faceted browsing is commonly used in web commerce to organize and visualize a set of products offered to buyers, based on the different attributes that products have: e.g. size, color, price, categories etc. Typically, if a web shop offers faceting by color, i.e. it displays information on all colors a product can have and the number of products available for each color. Selecting a color from the listing will refine the product selection to only products with the specified color. Thus, faceting serves multiple purposes. It presents the distributions of values for each property (dimension) and it allows for filtering by conditioning on the values of a particular dimension.

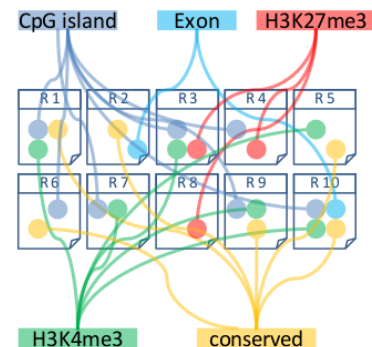
The most popular indexing structures for full text search is the *inverted index* (Zobel et al., 1998). The inverted index is a structure that is built from a set of text documents, by extracting all words appearing in the documents and for each word storing the documents in which the word appears. There is a popular extension of the inverted index called the *full inverted index*, where also the position of the word in the document is stored (Baeza-Yates and Ribeiro-Neto, 1999).

Among the popular search engines that are open source and publicly available we mention Solr (Apache Solr, 2004) and Elasticsearch (ElasticSearch, 2010). Solr and Elasticsearch provide efficient full text search with automated scaling and ranking and are both based on Lucene (Apache Lucene, 1999). The inverted index offers efficient indexing and instant query processing, however, it does not support efficient faceting and auto-completion. A novel indexing structure optimized for auto-completion, called HYB, was proposed by Bast and Weber (2006). HYB was shown to efficiently answer plain search queries, support faceting and instantly suggest automatic completions of partial queries. The HYB indexing was used as a base of a novel search engine called CompleteSearch (Bast and Weber, 2007). Through creative use of prefix indexing, CompleteSearch provides support for faceted navigation and query auto-completion and it has been shown to outperform more standard approaches based on inverted indices.

In this thesis, we used HYB indexing and CompleteSearch to enable interactive visual exploration of large epigenomic datasets using EpiExplorer.



### 3. Live exploration and global analysis of large epigenomic datasets using EpiExplorer



Representing epigenome  
annotations as words in text  
documents

The advancements of biological technologies enabled epigenome mapping consortia to generate resources of great value for studying epigenetic regulation (Bernstein et al., 2010; Consortium, 2004; Hudson et al., 2010). To maximize their utility and impact, bioinformatic tools are needed to facilitate the analysis of these data. In this chapter, we describe EpiExplorer, a methodology and software for exploring genome and epigenome data on a genomic scale. EpiExplorer proposes a novel method for visual exploration of large datasets that addresses a major challenge in epigenetics research using state-of-the-art information retrieval methods<sup>1</sup>. EpiExplorer is available at <http://epiexplorer.mpi-inf.mpg.de>

The chapter is organized as follows. First, we discuss in detail the novel concept and methods that make EpiExplorer useful to bioinformaticians and biologists with various bioinformatic backgrounds. We present the efficient and versatile text indexing scheme that allows EpiExplorer analyses to complete within seconds. We validate the approach by demonstrating how one can reproduce well established findings about the epigenetic properties of CpG islands by a short EpiExplorer session. Then, we describe an insightful analysis of the DNA hydroxymethylation dataset published by Szulwach et al. (2011) in

<sup>1</sup>The work presented in this chapter has been published in Halachev et al. (2012). I conceived the project with support from Christoph Bock and Thomas Lengauer. KH, TL and CB planned the research. KH and CB conducted the research. KH developed the methods and software. Felipe Albrecht extended the annotation database and assisted with software development. Hannah Bast contributed software, ideas and technical guidance. KH and CB wrote the manuscript of Halachev et al. (2012), parts of which are used in this chapter. Section 3.4.3 is based on a collaborative work with Laura Tološi published in Tološi et al. (2013).

relation to public reference genetic and epigenetic annotations. We show how with EpiExplorer we identify a small subset of biologically interesting 5-hydroxymethylated sites, that are suitable for further biological validation. In a separate study, we demonstrate how EpiExplorer can be used in a disease-specific study, as we identify epigenetic characteristics of regions corresponding to recurring DNA breakpoints in large cohorts of cancer tissues. Finally, we summarize the impact of EpiExplorer on the community during its first year after publication, by showing and discussing relevant user statistics.

### 3.1. Background

Understanding gene regulation is an important goal in biomedical research. Historically, much of what we know about regulatory mechanisms has been discovered by mechanism-focused studies on a small set of model genes (Mitchell and Tjian, 1989; Orkin, 1990). High-throughput genomic mapping technologies have recently emerged as a complementary approach (Hawkins et al., 2010); and large-scale community projects are now generating comprehensive maps of genetic and epigenetic regulation for the human and mouse genomes (Adams et al., 2012; Bernstein et al., 2010; Consortium, 2004; Satterlee et al., 2010). Substantial potential for discovery lies in better connecting mechanism-focused studies investigating the wealth of functional genomics and epigenomics data that are being generated. A handful of pilot studies highlight the value of combining high-throughput and mechanism-focused research (for example, in (Huarte et al., 2010; Mikkelsen et al., 2010; Musumuru et al., 2010)), but few research groups are equally proficient in bioinformatics, large-scale genomics and in-depth functional analysis to conduct highly integrated studies of gene regulation. A new generation of software tools could reduce those requirements by enabling user-friendly navigation and analysis of large genomic databases.

Genome browsers are probably the most widely known type of bioinformatics software related to genomics. They are currently the only software tools for navigating through genome data that are widely used, not only by bioinformaticians but also by biomedical researchers with little computational background. The strength of web tools such as the UCSC Genome Browser (Karolchik et al., 2011; Fujita et al., 2011; Karolchik et al., 2008), Ensembl (Birney et al., 2004; Flicek et al., 2010, 2008), Integrative genomics viewer (Robinson et al., 2011) and the WashU Human Epigenome Browser (Zhou et al., 2011) lies in their intuitive interface, which enables users to browse through the genome by representing it as a one-dimensional map with various annotation tracks. The extensive database of annotation tracks together with the intuitive, detailed and configurable visualization of individual gene loci have established genome browsers as tools of choice for many small-scale studies for exploring the properties of one or several loci. However, the focus on single locations is also the main limitation of genome browsers, as they lack the potential to perform truly genome-wide analysis of complex datasets by investigating multiple genomic regions together. Therefore, complementary tools are needed that suitably address the complexity of large genomic datasets while maintaining the interactive and user-friendly character of genome browsers.

Existing tools, however, do not fully address that need. For example, the UCSC Table Browser (Karolchik et al., 2004) and Ensembl BioMarts (Kinsella et al., 2011) provide user-friendly support for selecting and downloading sets of genomic regions. However,



the analysis of the downloaded data needs to be performed locally using command-line tools, such as BEDTools (Quinlan and Hall, 2010), BEDOPS (Neph et al., 2012) and R/Bioconductor (Gentleman et al., 2004) thus lacking an interactive and user-friendly interface. Workflow tools such as Galaxy (Goecks et al., 2010; Blankenberg et al., 2010), Taverna (Oinn et al., 2006; Hull et al., 2006) and the Genomic HyperBrowser (Sandve et al., 2010) combine user friendliness and flexibility, but they require careful planning and tend to be too slow for performing truly interactive and exploratory analyses. Finally, enrichment analysis servers such as GREAT (McLean et al., 2010) and EpiGRAPH (Bock et al., 2009) (see chapter 4) are powerful tools for identifying significant associations in large biological datasets, but they lack the flexibility to explore the observed enrichments in a dynamic and interactive fashion.

None of the tools discussed above enables the user to easily navigate through the plethora of publicly available genomic and epigenomic data and easily form hypotheses about the interplay of the various biological mechanisms. Similar challenges have been faced by a different branch of applied computer science over the last decade: the challenge of providing efficient and intuitive search in web content.

To provide efficient web search, search engines continuously process and index the available web content. During the last decade, web search engines have increased the number of web sites they index by several orders of magnitude, while maintaining almost instant search response. The algorithms used address the problems of providing intuitive and instantaneous exploration of large heterogeneous datasets, generated from various sources and following different formats. These properties also characterize genomic and epigenomic data. With proper adaptation, methods for indexing and searching in web content can also be of use for bioinformatic tools.

In Section 2.6.1, we presented information retrieval approaches for indexing and querying text data. Specifically, we discussed CompleteSearch (Bast and Weber, 2007), a search engine based on HYB data indexing (Bast and Weber, 2006) that enables not only standard text queries, but also prefix, range, negation and JOIN queries. Using CompleteSearch in an innovative manner, with biological data, we developed EpiExplorer, a new paradigm for live visual exploration of massive epigenetic datasets.

## 3.2. Concept and main features of EpiExplorer

With EpiExplorer, we developed a web server that combines the interactive nature of genome browsers with the region-based analytical approach of Galaxy. EpiExplorer is based on an efficient indexing structure powering the CompleteSearch engine that instantly answers search queries and provides faceted navigation. The service enables users to casually explore large-scale genomic datasets in search of interesting functional associations. EpiExplorer does not aim to replace any existing tool; instead it facilitates dynamic integration with tools such as the UCSC Genome Browser (Karolchik et al., 2008), Galaxy (Goecks et al., 2010) and the Genomic HyperBrowser (Sandve et al., 2010), it bases its preprocessing of a dataset on the efficient command-line tool BEDTools (Quinlan and Hall, 2010) and it uses the CompleteSearch text search engine (Bast and Weber, 2007) to provide instant responses. EpiExplorer does not expect the user to define a detailed framework for searching for relevant associations in the data — as enrichment analysis tools do with their statistical

testing environments. Instead, EpiExplorer’s key strength lies in supporting exploratory hypothesis generation using a broad range of genomic analyses performed in real time over the Internet. Such exploratory analyses often provide a first indication of relevant associations that are worth following up by in-depth statistical analysis using other software tools or by experimental validation in the wet lab.

EpiExplorer works with built-in or user-uploaded genomic region sets. Five types of genomic regions are available in EpiExplorer by default, namely CpG islands, gene promoters, transcription start sites, predicted enhancer elements and a map of 5-kb tiling regions spanning the entire genome. The user-uploaded datasets can be investigated with the same speed and flexibility as any of EpiExplorer’s default region sets. Every custom region set needs to be preprocessed to be available in EpiExplorer. During the dataset preprocessing phase, the set of genomic regions is annotated with various properties and is transformed into a data structure that allows efficient querying. More specifically, preprocessing involves the following steps:

1. The user prepares and supplies a list of genomic regions as a BED file.
2. The list of genomic regions is uploaded into the EpiExplorer server.
3. The genomic regions are internally annotated with a wide range of genomic and epigenomic attributes using BEDTools (Quinlan and Hall, 2010) and efficient scripting operations.
4. EpiExplorer represents each genomic region by a text document containing keywords for all its annotation features each region. Thus it creates a large virtual collection of text documents that represent the genomic regions and their properties.
5. CompleteSearch creates a search index for the collection of text documents representing the dataset.

After preprocessing, a user can load the dataset into the EpiExplorer interface and immediately start exploring it. At every step, EpiExplorer presents the user with visualizations that summarize different aspects of the dataset, as well as provide intuitive interface for follow-up steps, such as alternative visualizations and refinements. The power of EpiExplorer is the speed and ease with which the user can request and receive different visualizations and to refine the initial set of regions based on different filtering criteria. And despite our extensive reliance on text search for the above, the user never has to formulate any textual search phrases – they are dynamically constructed based on the user interaction with EpiExplorer’s graphical frontend.

EpiExplorer is based on four main concepts:

- *Mapping of biological data into text format.* The genomic region – epigenetic annotation is transformed into a document – word format, suitable for indexing and searching with efficient text-search algorithms. In order to utilize powerful text search operations for genomic analyses, we developed an encoding scheme that translates heterogeneous genome and epigenome datasets into a semi-structured text format. For that purpose, we designed a custom “dictionary” of keywords representing epigenetic properties; for example the keyword ‘overlap:CGI’ is present in documents

corresponding to regions that overlap with CpG islands. Moreover, keywords in EpiExplorer are structured hierarchically, which affords prefix search at various levels of granularity. For example, the term *overlap:histones:H3K4me3* selects all regions that overlap with an H3K4me3 peak in any tissue, while the more specific term *overlap:histones:H3K4me3:H1hESC* selects only those regions that overlap with an H3K4me3 peak in ES cells. Furthermore, we can perform auto-completion queries such as *overlap:histones:H3K4me3:\** that returns the number of regions that overlap with an H3K4me3 peak separately for each tissue. EpiExplorer also encodes various numeric scores (such as overlap ratios and DNA methylation levels) in a manner suitable for prefix (text) search.

- *Faceted visualizations.* EpiExplorer utilizes autocompletion queries supported by CompleteSearch to offer faceting visualizations on biological data. Faceting in web commerce serves multiple purpose. It offers easier navigation to specific products, but it also advertises and indirectly familiarizes online shoppers with different aspects of a product catalog. We transfer these concepts to biological data analysis by offering easy dataset refinements and intuitive visualizations, while advertising multiple other aspects of the datasets that are not subject to the current analysis. Overall, the use of the CompleteSearch engine for semi-structured text search confers a level of flexibility, efficiency and scalability that would not be easy to achieve with a simple text-tagging approach or with a relational database management system.
- *The real-time responsiveness of the interactive user interface.* Running an analysis with EpiExplorer means continuous interaction with the tool via its user interface. Through the user interface, the user can perform dynamic refinements and request custom visualizations and complex data views. All actions take place in real-time, enabling the user to focus on their dataset and not on the specifics of the tool.
- *Scalable software implementation.* EpiExplorer employs a scalable software infrastructure that allows for the analysis of thousands of custom datasets without the need of large-scale computational resources.

### 3.2.1. Functionalities of the EpiExplorer software server

In this section, we discuss the main functionalities of the EpiExplorer web service.

*Dynamic refinements.* EpiExplorer facilitates the refinement of a set of regions if provided a property the regions should or should not have. For example, if the current set of regions is 'gene promoters', the user can easily select only those gene promoters that overlap with 'CpG islands' to investigate the epigenetic properties associated with gene promoters and CpG islands. As a target for such refinements, the user can choose from hundreds of categorical and numerical properties (see Appendix A). When refining by a categorical property, the user is presented with a list of the possible values from which to select. When refining on numerical properties, the users is presented with visual sliders that enable to specify ranges of values that interest them. For example, when exploring the default dataset 'CpG islands', the lengths of the islands range from 200 basepairs to several thousand basepairs. The user can easily refine the dataset to only islands with lengths of at least 700bp, namely the longer (and assumed stronger) CpG islands.

*Interactive browsing of the properties of current region set via detailed faceted information.* One of the most prominent and frequently used features of EpiExplorer is the quick overview in the form of a column chart of the properties of a dynamic set of regions (see Figure 3.19 as an example). Queries like ‘how many regions from a specific set overlap with a certain genetic or epigenetic annotation’ are highly common in bioinformatic analyses and bioinformatic tools provide various approaches to answering them. Due to the appropriate data indexing, EpiExplorer is able to answer hundreds of such questions instantly. Via this functionality, the user can select a set of properties of interest and quickly get an overview on the profile of her set of regions with respect to these properties. For example, the user can find out what percentage of the regions overlap with CpG islands, with gene promoters, with several H3K4me3 and H3K27me3 histone modifications, with enhancers, with insulators, with repeats by directly from looking at the *summary* plot. Then with a single click she can follow up by inspecting the overlap with H3K4me3 in all available tissues. Finally, the exploratory power of EpiExplorer comes from the combination of dynamic refinements and instant visualization. Following dynamic refinement, EpiExplorer instantly updates the summary of properties for the new regions subset, thus enabling the user to explore practically any describable subset of the dataset, while progressively gaining insights into the biological makeup of the targeted regions.

*Custom dataset upload.* For each of the genome assemblies that EpiExplorer supports, it provides five standard, precomputed and ready-to-use datasets: CpG islands, gene promoters, gene transcription start sites, enhancers and tiling regions. These datasets serve a double purpose: to easily demonstrate to new users how EpiExplorer works and to easily check and confirm known and accepted biological results, for example that CpG islands are often overlapping with gene promoters, thus validating the EpiExplorer preprocessing. However, these precomputed datasets are mostly for introductory purposes. The interest of users lies mainly in exploring the properties of novel sets of regions. Users can upload their datasets to EpiExplorer and they are automatically annotated, indexed and made ready for investigation. The combination of instant browsing and dynamic refinements on custom datasets makes EpiExplorer useful to a wide range of biologists and bioinformaticians.

*Comparative analysis.* As the user investigates the properties of a set of regions, interesting biological observations (hypotheses) often arise. For example, a peak of a H3K4me3 histone modification occurring within a large percentage of the regions can either carry biological relevance or is expected by chance given the distribution of the peaks across the genome. In order to offer a rough estimate of the relevance of the association, EpiExplorer provides with a basic comparison to a randomized control set, which is automatically generated when custom datasets are uploaded into EpiExplorer. Randomized control sets are generated simply by reshuffling the genomic position of all regions in the dataset, in a procedure inspired by the popular permutation tests. If the user has a control set that suits the purpose of the analysis better, then that set can be provided to EpiExplorer as a reference set. Any analysis steps that are performed on the active set of regions (refinements), can either be simultaneously applied to the control set as well (*dynamic comparison*), or

the control set can remain unchanged (*static comparison*). EpiExplorer generates only one control set, which is not sufficient for running a proper permutation test and assessing statistical significance. More control sets would be necessary for variance estimation, however the computational effort might be too large and may lead to slowing down of the exploration of the data. Speed is one of the most important requirements of our web service, whereas statistical inference is not a priority. Instead, we facilitate direct export from EpiExplorer to external resources specialized in statistical inference such as EpiGRAPH (see Chapter 4) and the Genomic HyperBrowser (as illustrated in the corresponding tutorial on the supplementary website (EpiExplorer: supplementary information, 2012)). Despite its limitation, the comparison with a control set gives a fast visual intuition of the strength of the association between genomic attributes, association which has a high chance of being statistically significant.

*Flexible disjunctive(OR) refinements.* Successive refinements with EpiExplorer correspond to an conjunctive (*AND*) query. For example, if from a set of regions, the regions with property *P* are selected first and then from those only the regions with property *R* are selected, then all regions in the resulting set will have both the properties *P* and *R*. Combining of search terms with the *AND* operator is considered standard for search engines and is easily addressed algorithmically, whereas disjunctive (*OR*) queries are not always supported. The CompleteSearch engine supports *OR* queries. This feature extends the diverse filtering options for performing complex analyses. For example, the query `dnameth:ES:ratio:00–dnameth:ES:ratio:33 | dnameth:ES:ratio:66–dnameth:ES:ratio:99` selects all regions that have a methylation score between 0 and 0.33 or between 0.66 and 1. As methylation score ranges from 0 and 1, the filter above identifies only the strongly unmethylated and the strongly methylated regions in the process disregarding the regions with intermediate methylation.

*Sharing results.* We believe that reproducibility of bioinformatic analyses is extremely important, considering that the results they deliver are often tested in clinics and require large resources. EpiExplorer adheres to this requirement (Gentleman and Lang, 2004) by providing several ways of documenting and sharing analyses. Each user-uploaded region set is assigned a unique identifier that also serves as a password for accessing this dataset. Sharing this identifier with other researchers enables them to analyze the same dataset in EpiExplorer without any need for copying or transferring datasets. Furthermore, at any point in an EpiExplorer analysis, an identifying URL can be obtained to dynamically recreate the analysis and to enable the user to follow up on the results without affecting the original analysis snapshot. Because all steps of an EpiExplorer analysis are documented in the control menu, the snapshot functionality ensures that EpiExplorer analyses are readily reproducible (see tutorials on the supplementary website (EpiExplorer: supplementary information, 2012)). EpiExplorer also supports the export of its visualizations, of any region set as a downloadable BED file, as custom tracks in the UCSC Genome Browser and Ensembl, and the transfer into Galaxy and Galaxy-powered tools such as the Genomic HyperBrowser for further analysis; it also provides lists of gene identifiers for export to gene set tools such as DAVID and Gene Set Enrichment Analysis (GSEA).

*Data privacy.* Every custom dataset, refinement and visualization is accessible only to its creator (unless explicitly shared with other researchers) and protected by unique identifiers functioning as passwords, thus ensuring the privacy of data and analyses. When a user uploads a dataset, the data is sent directly to an EpiExplorer server behind a firewall. No part of the data is ever stored on the EpiExplorer web server and thus is never directly accessible from the web. After preprocessing of the custom dataset the user is provided a unique identifier. The identifiers are formed by the dataset name provided by the user coupled with a randomly generated string that makes it impossible to access the data of other users by chance. The only way to access a dataset is via its identifier, and initially the identifier is known only by the person uploading the dataset. We also provide a manual service to fully erase all user data on the EpiExplorer servers. For the purpose, the user needs to send us an email, from the email associated with the dataset, specifying the dataset identifier and explicitly requesting the deletion of the dataset.

### 3.3. Methods

In the following section, we present the details of the EpiExplorer method and software. We start by introducing the textual encoding scheme EpiExplorer uses to store and efficiently query epigenetic and genetic properties. Then, we discuss how custom datasets are preprocessed. We continue by presenting the software architecture and implementation details followed by an overview of the common workflows when using EpiExplorer. We conclude with a detailed presentation of the user interface elements as well as descriptions of the main operations.

#### 3.3.1. Translating biological concepts to text

EpiExplorer internally represents each genomic region as a text file that encodes region-specific annotations in a semi-structured text format. For binary and categorical attributes (such as a region's association with an H3K4me1 peak or a 5hmC hotspot), the key concept is *overlap*. Two genomic regions are treated as overlapping if they have at least one base pair in common, and it is often plausible to assume that region sets that overlap more frequently than expected by chance are involved in similar biological processes (for example, co-binding of functionally related transcription factors). EpiExplorer precomputes two additional overlap concepts: *overlap10%* and *overlap50%*. A dataset region is considered to *overlap50%* with an annotation if at least 50% of the region is covered by regions from the annotation. To effectively handle such data in the context of text search, we define the prefix *overlap:* followed by an annotation identifier. For example, the word *overlap:genes* indicates that the current region overlaps with the body of a gene, *overlap:conserved* encodes the overlap with a conserved element, and *overlap:CGI* denotes overlap with a CpG island. Using CompleteSearch's prefix search functionality, it we can efficiently retrieve all completions of a given prefix. For example, the query *overlap:\** retrieves all possible completions of the prefix *overlap:*, reporting the number of regions for each completion (see Figure 3.10 for an example). In this way, overlap information for a large number of genome and epigenome attributes can be obtained via a single text search query that is almost always answered within seconds (see Table 3.2 from Section 3.4.5 for details). Fur-

thermore, the general overlap query *overlap:\** can be refined according to the hierarchical structure of the encoding scheme. For example, the query *overlap:histones:H3K4me3:\** retrieves an overlap summary of the H3K4me3 mark for all cell types included in EpiExplorer, whereas *overlap:histones:H3K4me3:ES* only obtains the regions that overlap with an H3K4me3 peak in ES cells.

### Textual encoding of binary and categorical genomic attributes

In order to utilize text search algorithms for interactive exploration of large biological datasets, we developed an encoding scheme that translates heterogeneous genome and epigenome datasets into a semi-structured text format. Each genomic region (e.g. CpG island or 5hmC hotspot) is represented by a text document containing keywords for all its annotation features (see Section 3.3.2). This textual encoding utilizes a prefix format that readily supports binary and categorical attributes (e.g. as overlap with H3K4me1 peaks or association with a Gene Ontology term) as well as numerical attributes (e.g. region length or 5hmC hotspot significance). The resulting collection of text documents — each representing a genomic region with extensive genome and epigenome annotations — can be searched in a highly efficient manner using the CompleteSearch engine (Bast and Weber, 2007). CompleteSearch implements an index structure that was specifically designed for feature-rich search in semi-structured text (Bast and Weber, 2006). Through creative use of prefix indexing, CompleteSearch provides support for a number of advanced features such as query autocompletion, semi-structured text search and database-style JOIN operations (Bast and Weber, 2007). EpiExplorer makes use of these features in order to implement complex operations in a highly efficient manner. For example, we combine the autocompletion feature with a hierarchical encoding scheme for genomic annotations in order to produce each of EpiExplorer’s diagram types with a single query to the CompleteSearch engine. Similarly, we utilize JOIN operations to perform complex refinement operations that combine region-based filtering with gene-based filtering. The empirical performance of EpiExplorer is summarized in Table 3.2 in Section 3.4.4. For example, more than 99% of approximately 4,000 queries that have been run on the user-uploaded 5hmC hotspot dataset consisting of 82,221 genomic regions completed in less than two seconds. EpiExplorer runs a dedicated instance of the CompleteSearch engine for each set of genomic regions (see Figure 3.6), which makes the software highly parallel and scalable to very large numbers of user-defined region sets. Figures 3.7 to 3.10 illustrate the typical workflow of an EpiExplorer analysis. Once a user-defined region set has been uploaded, the middleware annotates each region with data from EpiExplorer’s genome and epigenome annotation database, encodes these annotations as structured text and creates a dedicated CompleteSearch instance supporting search on this region set. For every analysis that is requested via the user interface, EpiExplorer’s middleware constructs a text search query that is then sent to the corresponding CompleteSearch instance. The text search engine runs the query against its precalculated index and returns a set of matching regions. The middleware decodes the textual format and passes the results on to the user interface. Although EpiExplorer is internally implemented as a text search engine for genomic datasets, the text search is not visible by the user. Search results are automatically converted into visual representations, which have been designed to facilitate the discovery

of relevant biological associations and the identification of strong candidates for follow-up research. EpiExplorer also provides a straightforward way for users to keep track of their results and to share specific datasets or analyses with their collaborators. When a custom dataset is uploaded, EpiExplorer assigns to it a unique identifier URL, which also serves as a password for accessing this dataset. Sharing this identifier URL with collaborators allows them access to the dataset. Analogously, an identifier URL can be obtained at any point during an EpiExplorer analysis, providing a permanent snapshot of the dataset, analysis and filtering steps that led to the result. When the snapshot-identifying URL is shared with other researchers, it enables them not only to reproduce the analysis but also to edit, customize and refine it independently of the original snapshot. These features make EpiExplorer well-suited for data sharing and reproducible research while providing strong data protection for user-specific analyses and datasets.

### Textual encoding of numeric genomic attributes

Many genomic attributes are numeric - for example, the CpG content or the distance to a neighboring gene. To be able to perform efficient text search on these attributes, we limit their numerical precision (number of digits) to a fixed number and use a binning scheme when necessary. We can then incorporate numeric score values into the textual encoding scheme by creating words such as *dnaseq:freq:CG:010*, which indicates that a genomic region exhibits a CpG frequency of 0.010 (1.0%). This textual encoding enables EpiExplorer to retrieve the distribution of CpG frequencies in a set of regions using the prefix query *dnaseq:freq:CG:\**, which facilitates efficient data collection and plotting of histograms. Using CompleteSearch's range query feature, it is also straightforward to obtain all genomic regions with numeric attributes that fall into a certain range. For example, the query *dnaseq:freq:CG:010-dnaseq:freq:CG:050* retrieves only those regions that have a CpG frequency of at least 1% and not more than 5%. Beyond region score attributes, additional numeric attributes supported by EpiExplorer include overlap ratios for filtering on the percent overlap between genomic regions as well as distances to neighboring genomic elements, which enable filtering steps such as 'identify all regions within 20 kb from the nearest gene'. Binary, categorical and numeric queries can be combined and iteratively refined in arbitrary ways. For example, the query *overlap:CGI dnaseq:freq:CG:010-dnaseq:freq:CG:050* retrieves all regions that overlap with CpG islands and exhibit a CpG frequency in the range of 1% to 5%.

### Integration of gene-centric textual annotations

In addition to binary, categorical and numeric attributes, EpiExplorer also incorporates textual information that is associated with genes, which includes Gene Ontology terms and OMIM phenotypes. As these annotations are already in text format, they can be used directly as keywords in the text search index. However, because these textual annotations can be lengthy and often apply to multiple genomic regions overlapping with the same gene, it is not ideal to store them directly in the description of each region. Instead, EpiExplorer maintains genes and their textual annotations as separate documents and stores only the gene identifier in the annotation of every overlapping genomic region. For example, if a region overlaps with the BRCA2 gene, EpiExplorer will add the word *gene:BRCA2* to



the document that represents the region, while the lengthy textual annotations of *BRCA2* are stored in a separate document named *gene:BRCA2*. To answer text search queries that include these gene annotations, EpiExplorer makes use of the database *JOIN* feature that is supported by CompleteSearch. In this way, the results from a region-based search and the results from a gene-based search can be combined in a single query, and only the matches are returned for visualization.

### 3.3.2. Precomputing epigenomic and genomic properties of genomic regions

The main steps that prepare a dataset for analysis with EpiExplorer are:

1. the user uploads a set of regions,
2. EpiExplorer annotates each genomic region with properties from its database,
3. EpiExplorer creates a data index of the dataset regions and their properties and provides the user with the unique identifier of the dataset.

First, a dataset in a standard BED format is sent to EpiExplorer for processing (Figure 3.1). EpiExplorer requires that every region is specified by its chromosome and the chromosome start and end coordinates. Additionally, EpiExplorer offers to utilize strand information if such is available. Moreover, the user may provide an additional score column, which contains some quantitative indicator relevant for the analysis, that can be used for interactive refining with EpiExplorer. For example, the BED file given by the user may contain a column with a score for each region, which indicates the strength of the property. The score is encoded into EpiGRAPH and indexed, thereby the user is able to easily refine the regions to subsets of certain strength, or for example to compare strong and weak regions.

Step	Description	Representation			
1. Upload	The user uploads a set of genomic regions (in standard BED format)	region	chrom	start	end
		Region 1	chr1	1000	4240
		Region 2	chr2	500	1545
		Region 3	chr1	8300	8850
		Region 4	chr5	3100	3400

Figure 3.1.: Genomic region files are uploaded to EpiExplorer in the UCSC Genome Browser’s BED format (UCSC Genome Browser BED format documentation, 2011)], with mandatory columns specifying the chromosome, start and end positions of each region

As a next step, every genomic region is annotated with multiple properties (Figure 3.2), such as *length*, *frequency of DNA patterns*, *overlap with CGI*, *distance to nearest CGI* and many, many others. The EpiExplorer backend uses the BEDTools (Quinlan and Hall, 2010) software to compute some of the properties such as overlap percentages and distances to the nearest genomic region from an annotation. By default, EpiExplorer computes all

Step	Description	Representation					
2. Annotate	Each genomic region is annotated with a broad range of genomic attributes	Region	chrom	Length	Frequency of CpG	Overlap with CGI	Distance to nearest CGI
		Region 1	chr1	3240	0.07	34%	0
		Region 2	chr2	1045	0.02	0%	521
		Region 3	chr1	550	0.05	5%	0
		Region 4	chr5	300	0.16	80%	0

Figure 3.2.: The EpiExplorer middleware annotates the uploaded regions with qualitative and quantitative attributes such as overlap with CpG islands and distance to the nearest CpG island

available properties of the custom datasets, but it also offers the possibility of choosing a custom subset of properties, thus significantly speeding up the preprocessing time.

Once the relevant qualitative and quantitative attributes of each region are computed, EpiExplorer converts these data into a collection of text documents (Figure 3.3). This is done by creating a separate text document corresponding to each genomic region and adding to it the words that correspond to the computed properties for this region. Previously in this section, we discussed the details of the textual encoding of the various attributes (see Section 3.3.1).

Step	Description	Representation			
3. Convert to text	Every region is represented as a text document and its annotations are translated into words	Region 1	Region 2	Region 3	Region 4
		chr1 length:3240 frequency:CG:07 overlapratio:CGI:34 overlap:CGI	chr2 length:1045 frequency:CG:02 distanceTo:CGI:521	chr1 length:0550 frequency:CG:05 overlapratio:CGI:05 overlap:CGI	chr5 length:0300 frequency:CG:16 overlapratio:CGI:80 overlap:CGI

Figure 3.3.: A text document is created for every genomic region, and its annotations are encoded in a semi-structured text format

Next, the CompleteSearch indexing scheme receives the collection of documents and starts creating an efficient HYB index. For that purpose, first two sorted lists are created, the list of all occurring words sorted alphabetically and the list of all documents sorted by document ID (Figure 3.4).

Finally, CompleteSearch splits the sorted list of words into ranges. Each range is stored efficiently into a single memory block. The advantage of the memory blocks is that it enables very efficient prefix search within a block. Each blocks stores in a sorted manner all pairs (wordId, documentId) for which the wordId is within the word range corresponding to the block and documentId is a document in which the word wordId appears (Figure 3.5).

Step	Description	Representation					
4. Sort	Words and documents are sorted & assigned unique identifiers	Doc ID	Document	Word ID	Word	Word ID	Word
		D1	Region 1	W1	chr1	W9	length:0300
		D2	Region 2	W2	chr2	W10	length:0550
		D3	Region 3	W3	chr5	W11	length:1045
		D4	Region 4	W4	distanceTo:CGI:521	W12	length: 3240
				W5	frequency:CG:02	W13	overlap:CGI
				W6	frequency:CG:05	W14	overlpratio:CGI:05
				W7	frequency:CG:07	W15	overlpratio:CGI:34
				W8	frequency:CG:16	W16	overlpratio:CGI:80

Figure 3.4.: To enable alphanumeric search for numeric attributes, leading zeros are added until all numbers have the same number of digits. Two sorted lists are created, one containing document identifiers and the other containing keyword identifiers

Step	Description	Representation			
5. Create index	Sorted lists are stored in memory such that blocks correspond to ranges of word IDs and contain all pairs of document/word IDs in a given range	Block	Word ID range	Corresponding words	document-word pairs
		B1	W1 - W3	chr1, chr2, chr5	(D1,W1) (D2,W2) (D3,W3) (D4,W1)
		B2	W4 - W8	distanceTo:CGI:521, frequency:CG:02, frequency:CG:05, frequency:CG:07, frequency:CG:16	(D1,W7) (D2,W4) (D2,W5) (D3,W6) (D4,W8)
		B3	W9 - W12	length:0300, length:0550, length:1045, length:3240	(D1,W13) (D1,W15) (D3,W13) (D3,W14) (D4,W13) (D4,W16)
		B4	W13- W16	overlap:CGI, overlpratio:CGI:05, overlpratio:CGI:34, overlpratio:CGI:80	(D1,W13) (D1,W15) (D3,W13) (D3,W14) (D4,W13) (D4,W16)

Figure 3.5.: CompleteSearch creates a text index connecting sorted word identifier ranges with sorted document-word pairs, which handles text search queries in a highly efficient fashion .

### Genomic and epigenomic annotations of region sets

EpiExplorer makes no conceptual distinction between default and user-uploaded region sets. Every feature that is available for default region sets can also be used on custom data. Upon upload, new region sets are automatically annotated with a broad range of genome and epigenome attributes that are maintained in EpiExplorer's annotation database (see Appendix A). The user can also select custom region sets as annotations for other user-uploaded region sets. The current version of EpiExplorer provides full support for the human genome assemblies hg18/NCBI36 and hg19/GRCh37, as well as for the mouse genome assembly mm9/NCBIM37. By default, EpiExplorer annotates every region with its chromosomal position, region length, strand and score attributes (if included in the uploaded BED file), and with annotations of ten different types: DNA sequence composition, histone modifications, transcription factor binding sites, DNaseI hypersensitive sites, DNA methylation, chromatin state segmentation, CpG islands, evolutionary conservation, repeat

elements and gene-associated attributes. These annotations are derived from the following sources: (i) DNA sequence composition attributes are calculated directly from the genomic DNA sequence, which was downloaded from the UCSC Genome Browser (Karolchik et al., 2008). (ii) Histone modification data have been generated as part of the ENCODE project (Consortium, 2004) and were obtained from the UCSC Genome Browser (Raney et al., 2011). We used preprocessed peak regions for 11 histone modifications and chromatin marks (H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me1, H3K27ac, H3K27me3, H3K36me3, H4K20me1, CTCF and Pol2) in nine cell lines (GM12878, H1hESC, HepG2, HMEC, HSMM, HUVEC, K562, NHEK and NHLF; described in more detail in the ENCODE documentation (ENCODE Common Cell Types, 2013). (iii) Experimental data for transcription factor binding have also been generated as part of the ENCODE project and were obtained from the UCSC Genome Browser. We used preprocessed peaks for 33 transcription factors (AP2alpha, AP2gamma, ATF3, BDP1, BRF1, BRF2, cFos, cJun, cMyc, E2F1, E2F4, E2F6, GATA1, GATA2, GTF2B, HELFe, junD, MAX, NFE2, NFKB, Pol2, Pol3, Rad21, RPC155, SETDB1, SIRT6, TFIIC110, TR4, XRCC4, YY1, ZNF263, ZNF274 and ZZZ3) in at least one cell line. (iv) DNA methylation data have been generated and preprocessed in the context of the Roadmap Epigenomics initiative (Human Epigenome Atlas, 2013) as described previously (Bock et al., 2010a; Gu et al., 2010). They include ten tissue types: ES cells, fetal brain, fetal heart, fetal kidney, fetal lung, fibroblasts, hematopoietic progenitor cells, skeletal muscle, smooth muscle and stomach mucosa. (v) Chromatin segmentation data were obtained from a recent paper describing a hidden Markov model segmentation of histone modification data from the ENCODE project (Ernst et al., 2011). (vi) DNaseI hypersensitive sites were also obtained from the ENCODE project. (vii) CpG island annotations were downloaded from the UCSC Genome Browser ('CpG islands (specific)') and from the CgiHunter website ('CpG islands (sensitive)') (CgiHunter, 2013). (viii) Evolutionary conservation data were obtained from the phastCons annotation track of the UCSC Genome Browser (Siepel et al., 2005). (ix) Repeat element annotations were obtained from the RepeatMasker annotation track in the UCSC Genome Browser (Smit et al., 2010). (x) Gene-associated attributes were retrieved via Ensembl Biomart (Kasprzyk et al., 2004) and include the gene name, textual description as well as annotations from the Gene Ontology (Ashburner et al., 2000) and OMIM (Hamosh et al., 2005) databases.

### 3.3.3. Software architecture

EpiExplorer is implemented according to a three-tier architecture scheme (Figure 3.6). The web-based user interface communicates with EpiExplorer's middleware, which in turn is supported by an annotation database and dynamically loaded text search engines in the backend. The web-based interface enables users to explore, upload and refine genomic region datasets. The interface is highly dynamic through the combination of server-side scripting (in PHP) and client-side scripting (in JavaScript). EpiExplorer utilizes the jQuery library (jQuery, 2013) for implementing flexible client-side interface functionality and Google Chart Tools (Google Chart Tools, 2013) for generating interactive visualizations of the data. (The charts used by EpiExplorer do not exchange any data with Google or other servers and therefore do not compromise data privacy in any way.) All visual-

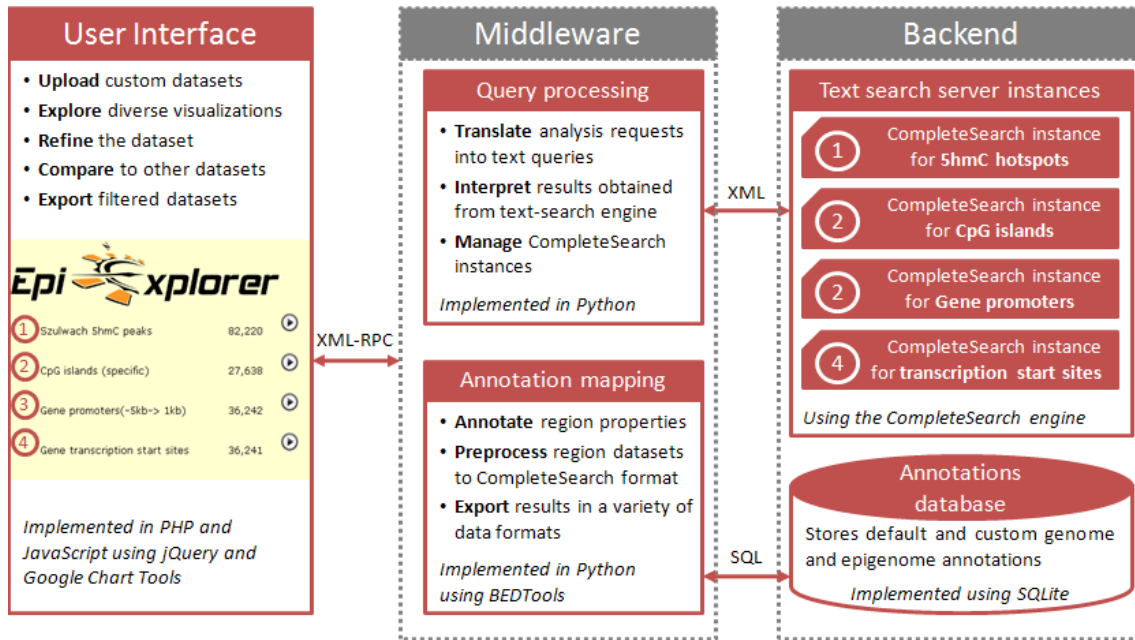


Figure 3.6.: Schematic outline of EpiExplorer’s software architecture, consisting of a web-based user interface, a query-processing and annotation-mapping middleware, and a text-search backend. The user interface is a dynamic web-based frontend implemented in PHP and JavaScript. The EpiExplorer middleware layer is implemented in the Python programming language and has two separate components: the query processing and annotation mapping. The backend of EpiExplorer consists of an annotation database implemented using SQLite and a collection of CompleteSearch server instances (one for each region set) that respond to text search queries sent by the middleware.

izations are dynamically generated based on region set data obtained via an XML-RPC connection with the middleware. The EpiExplorer middleware layer is implemented in the Python programming language and has two separate components. First, the annotation mapping module uses BEDTools (Quinlan and Hall, 2010) in combination with an annotation database (in the backend) to annotate user-uploaded datasets with genome and epigenome data. These annotations are translated into a semi-structured text format, and then used by the CompleteSearch index builder (Bast and Weber, 2007) to create a text index and a CompleteSearch instance corresponding to the dataset. Second, the middleware’s query processing module receives analysis requests from the web frontend, translates them into text search queries and polls the CompleteSearch instance that hosts the corresponding genomic region set. The CompleteSearch engine returns the results to the middleware, which decodes the text format and sends the results back to the user interface for visualization. The process also handles the active CompleteSearch instances. The backend of EpiExplorer consists of an annotation database implemented using SQLite and a collection of CompleteSearch server instances (one for each region set) that respond to text search queries sent by the middleware. The backend can be parallelized across multiple servers to increase performance. Unused CompleteSearch instances are automatically

suspended to disk by the middleware query server, from where they can be reactivated with minimal delay.

### **Scalability of the EpiExplorer software implementation as user load or computational demand increases**

To be able to handle the wave of epigenome data produced by international consortia, EpiExplorer was designed to scale to high user load and to be readily extensible with additional datasets. Because of the parallel nature of the computation-heavy backend, performance bottlenecks resulting from increasing user load can be resolved simply by adding more compute nodes for the backend. Furthermore, due to dynamic loading of backend instances, only parts of the indices of those region sets that are actively used need to be kept in memory, while additional user datasets are quickly reloaded from hard disk when a user accesses them. In its current version, EpiExplorer already handles hundreds of genome and epigenome annotations (Appendix A) and hundreds of custom datasets, even though we are not currently utilizing all the parallelization options that the EpiExplorer architecture provides. For example, during its first year as a public service, EpiExplorer has been working on a single machine that handles the preprocessing and user requests to thousands of custom datasets. As the demand increases these can be parallelized easily on multiple compute nodes.

### **Extensibility of the EpiExplorer concept by adding new annotation datasets, genome assemblies, novel data and analysis types**

Incorporating new datasets into EpiExplorer is straightforward and can be done by any user, provided that the data are available in (or can be converted to) one of several supported data types, namely genomic regions with or without a quantitative score and optionally including additional annotations such as strand information. For example, adding a new histone modification requires just a few mouse clicks in the frontend and less than an hour of computation time for the middleware and backend. Adding support for new genome assemblies is also relatively straightforward though not fully automated, as it requires minor modifications of the frontend and middleware. We demonstrated this as we integrated in EpiExplorer the TFBS and histone data from mouse ENCODE data (Mouse ENCODE Consortium et al., 2012) within a week of its publication<sup>2</sup>. Finally, the textual encoding behind EpiExplorer is flexible enough to incorporate conceptually new data types (for example, three-dimensional genomic interaction maps that link two or more genomic regions together), which would require modifications in the middleware's annotation mapping component and the implementation of new diagram types (for example, Circos plots (Krzywinski et al., 2009)) in the frontend. The source code of EpiExplorer is freely available for download from the support menu on EpiExplorer's supplementary website (EpiExplorer: supplementary information, 2012).

---

<sup>2</sup>Together with Felipe Albrecht

### 3.3.4. Computation workflow when processing a typical EpiExplorer query

For every data-analysis action that the user performs via the user interface, EpiExplorer's middleware constructs a text search query that is sent to the corresponding CompleteSearch instance. The text search engine runs the query against its index and returns a set of matching regions. The middleware decodes the textual format and passes the results on to the user interface, which visualizes the data in ways that facilitate intuitive exploration of genomic datasets. This computational approach makes it possible to solve complex non-textual analysis problems using single queries to a text search index, and thereby it enables the real-time exploration of large genomic datasets.

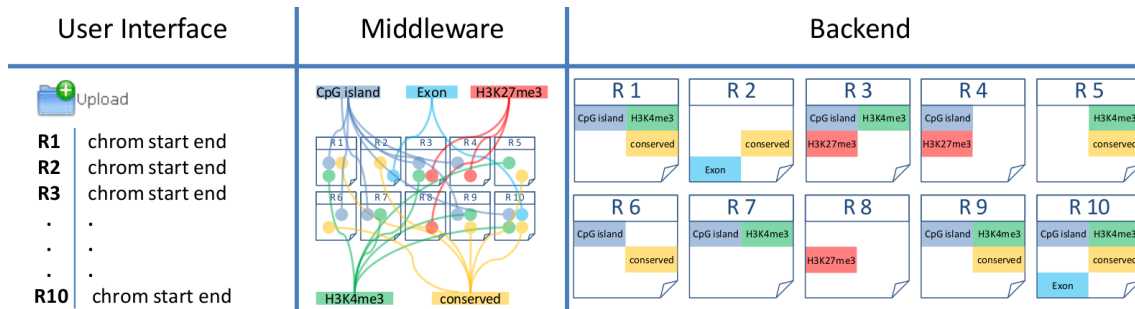


Figure 3.7.: When a user uploads a genomic region set (here: chromosome, start and end position for ten regions named R1 to R10), the middleware annotates this region set with genome and epigenome data, encodes the results in a semi-structured text format, and launches a CompleteSearch server instance to host the corresponding search index.

Figure 3.8 shows how EpiExplorer processes the request to retrieve all regions that have the property that they *overlap* with a CpG island. In this example, during the preprocessing phase every region that overlaps with a CpG island is marked with the artificial word *overlap:CGI*. Hence, when the user indicates his request the user interface sends the query *overlap:CGI* that is matched by CompleteSearch against the index of all regions. CompleteSearch then retrieves the number of regions that contain the word *overlap:CGI* and if indicated in the query returns the list of the regions.

Figure 3.9 shows how EpiExplorer processes requests with more than one term. In this case, the user request all regions that both overlap with a CpG island and also co-localize with an H3K4me3 peak. Again, there are artificial words that correspond to both properties. As a consequence, the query that retrieves the regions with the requested properties combines the two terms: *overlap:CGI overlap:H3K4me3*. The subset of regions that contain both properties are sent back to the user interface and visualized accordingly.

Finally, Figure 3.10 demonstrates one of the most innovative features of EpiExplorer. With a single prefix query, detailed faceted information for a region set is retrieved. In the examples above, we already pointed out that the textual representation of the regions contains multiple words like: *overlap:CGI*, *overlap:H3K4me3*, *overlap:exons* etc. The common characteristic of these words is that all of them are *overlap* properties, being encoded with the same prefix –"*overlap:*". Thus if CompleteSearch is presented with the prefix query "*overlap:\**", it runs all possible completions of the prefix. For each completion, the

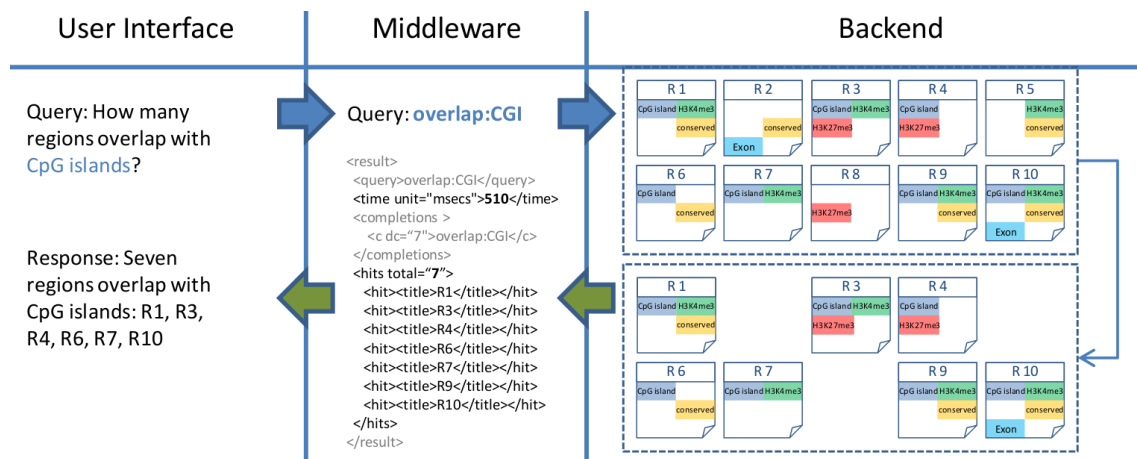


Figure 3.8.: To identify which regions overlap with a CpG island, a simple query overlap:CGI is sent to the backend, and the backend returns an XML file with the matching regions.

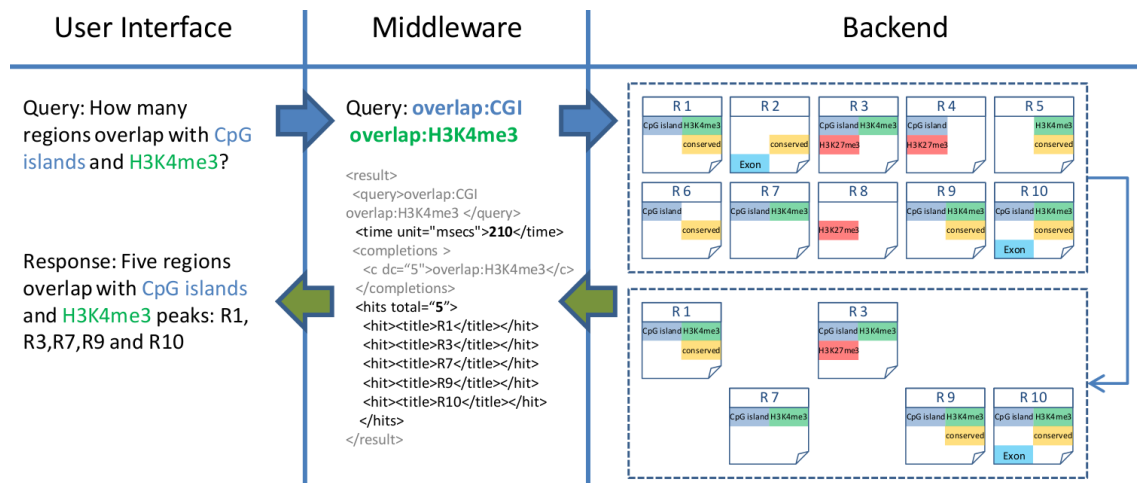


Figure 3.9.: To identify regions that overlap with CpG islands as well as with H3K4me3 peaks, an AND search is performed (query: overlap:CGI overlap:H3K4me3), and the backend returns only regions that are annotated with both keywords.

exact number of documents that contains the term is returned, which is the quantitative information that EpiExplorer uses for visualization. Thus, the result of the prefix query is an informative bar chart that summarizes the genomic and epigenomic context of the inspected set of regions.



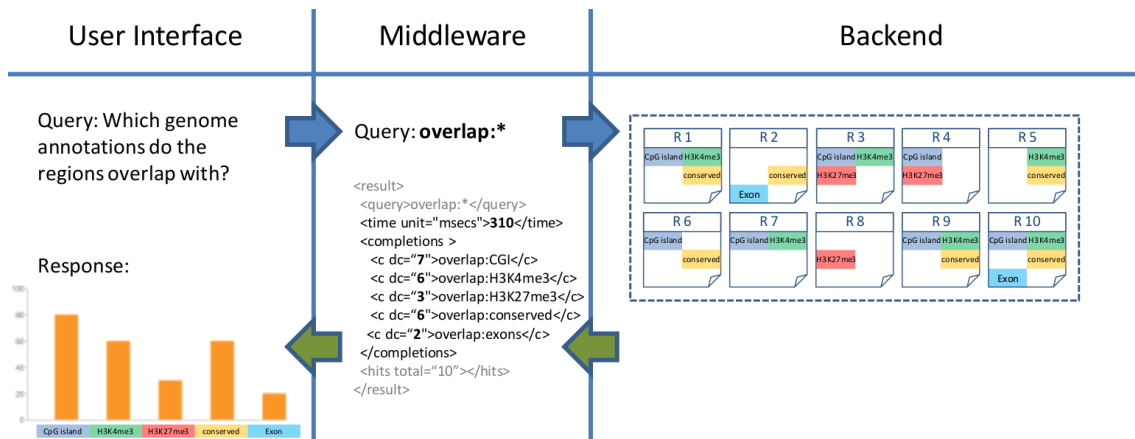


Figure 3.10.: To efficiently generate percent overlap diagrams, a prefix query `overlap:*` is sent to the backend, which identifies all possible completions of the prefix and returns the total number of regions matching each query completion.

### 3.3.5. EpiExplorer user interface and user experience

The EpiExplorer backend is complemented by a visual and intuitive frontend. In this section, we discuss the EpiExplorer user interface elements and how they provide dynamic and informative visualizations.

#### Basic elements of the user interface

When starting to work with EpiExplorer, the user needs to select a dataset to explore. The interface of EpiExplorer shows a screen that facilitates data input. Figure 3.11 presents the dataset selection view.

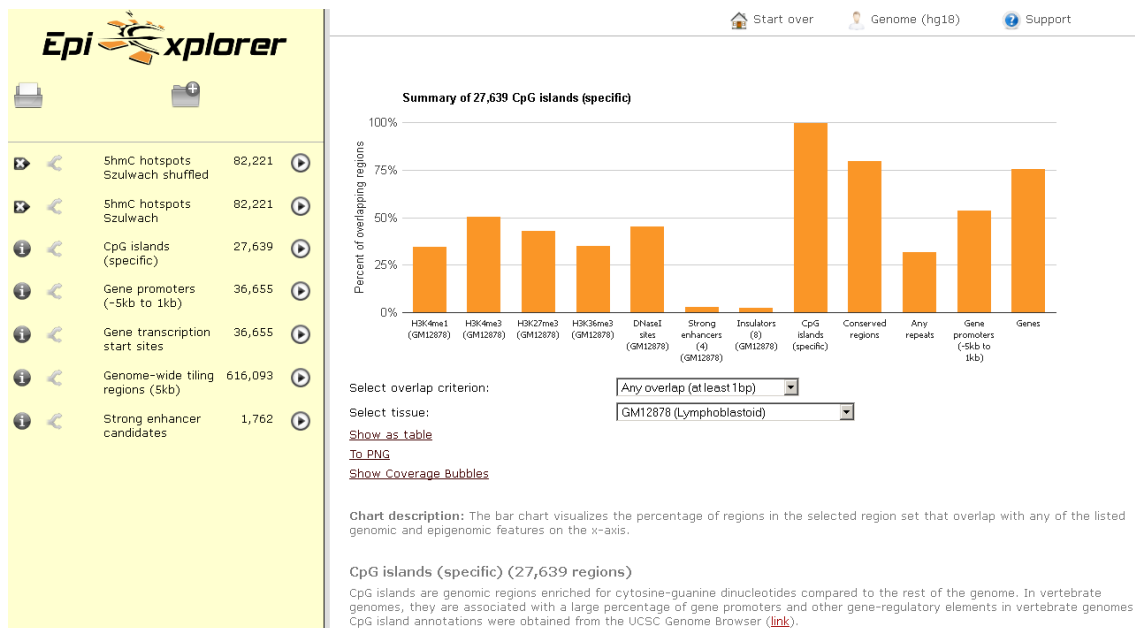


Figure 3.11.: Selecting a dataset view of EpiExplorer

The right-hand side of the screen is split into two sections, the top part contains a visualization and the bottom part contains explanatory text about the dataset and about the visualization. On the left-hand side of the screen, EpiExplorer shows the currently available datasets as well as the two buttons that can be used to introduce new datasets, the ‘Reload a dataset’ button in the top left and ‘Upload a new custom dataset’ next to it. Below them, the user sees the custom datasets that have been already loaded and below them the default datasets provided by EpiExplorer. For every dataset, we show the dataset name and the number of regions. To the left of the dataset name there are two buttons. For a custom dataset, the user can remove it from his list from there. The other button activates the comparison mode, that is explained in more details later (see Section 3.3.5). Clicking on each dataset name or on the button to its right selects the dataset for exploration and shows the exploration mode as shown in Figure 3.12. In exploration

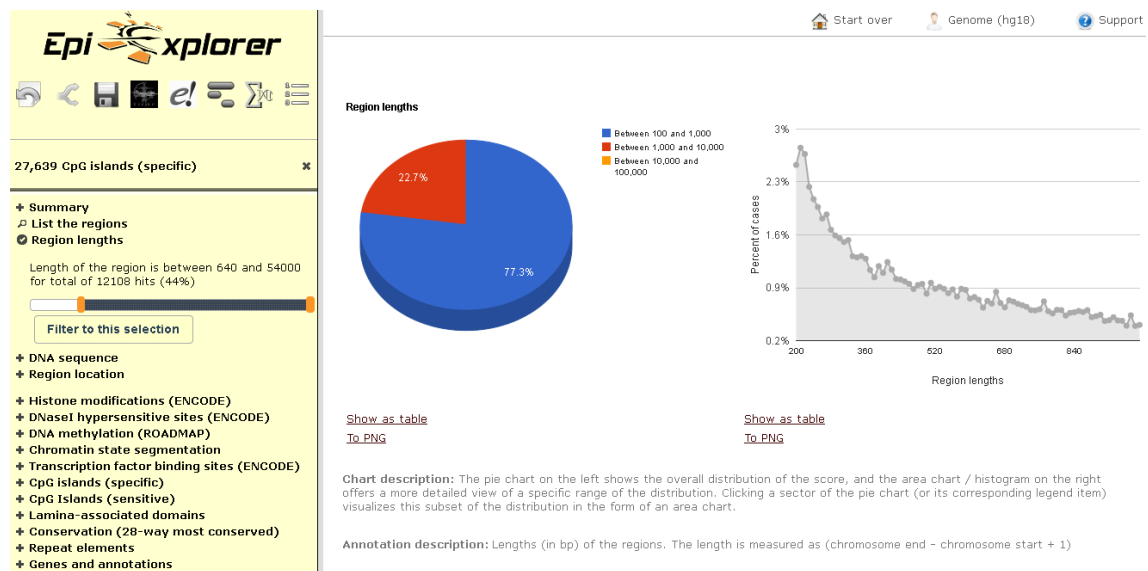


Figure 3.12.: Exploring a dataset view in EpiExplorer

mode, the user interface is split into three main parts. The top of the right-hand side shows visualizations and their settings and the bottom of the right-hand side shows information about the visualization, annotations used to compute it and details about the methods (if available). The left part of the screen displays possible actions. On top is the management menu, below are the current dataset and all applied refinements. Further below them is a listing of the possible annotations and the different visualizations and refinements that can be performed for each. When a user selects a visualization (for example, in Figure 3.12 the regions lengths are selected), then the right-hand side of the screen updates automatically with the relevant visualization and additional information. Also on the left-hand, in the yellow region a control appears that enables the user to refine the current selection. In this example, the user can use the controls to select a custom range of regions lengths; as the control is moved the information just above it changes dynamically to inform on how many regions (and what percentage) of the current selection have values in the selected range. If the user selects the "Filter to this selection" button, the current selection is instantly refined, the visualization on the left is updated to reflect the new subset of regions; any

analysis performed from this point on will only be applied on this selection. In Figure 3.13, we show the selection that results from applying to all CpG islands two refinements: overlap with a H3K4me3 peak in H1hESC tissue and overlap with a gene promoter regions. The listing indicates that there are 13,106 such regions identified. The controls on the right allow to remove a single refinement, to add an *OR* clause to a refinement or to completely reset the dataset selection.

<b>13,106 CpG islands (specific)</b>		<b>×</b>
The region overlaps with H3K4me3 sites in H1hESC tissue(1bp)	↔	<b>×</b>
The region overlaps with gene promoters (-5kb->1kb)(1bp)	↔	<b>×</b>

Figure 3.13.: Refining a selection of an EpiExplorer dataset

Finally, the management menu (see top left of Figure 3.12) contains quick links to some of the popular features. The first button resets all dataset refinements and puts the program back into dataset selection mode. The second button activates the comparison mode. The third button creates a URL link to the full state of the current analysis, including dataset selection, added refinements and chosen visualization. This functionality enables users to easily store an exact analysis state. The next four buttons allow the direct export of the current selection to four popular tools: UCSC Genome Browser, Ensembl, Galaxy and the Genomic HyperBrowser. The last button, moves the program into the export mode from which the user can export the full list of genomic regions in the current selection

EpiExplorer packs a lot of functionalities that allow to dynamically refine datasets, change visualizations and reset selection with a single click. This can be overwhelming for some first-time users. For that purpose, we introduced dynamic notifications that inform the user whenever he performs a specific action for a first time. The notifications explain the consequences of the action and the reason why it was taken. When the user feels confident in his understanding of EpiExplorer, he can easily switch off these notifications.

### EpiExplorer visualizes analysis results using six types of dynamically generated diagrams

The *bar* chart (see Figure 3.28 for an example) reports the percentage overlap of a selected region set with genomic regions of different types. Using the EpiExplorer control menu, it is straightforward to restrict a region set to those regions that overlap (or do not overlap) with another type of genomic regions shown in this diagram.

The *area* chart (see Figure 3.32 for an example) is essentially a histogram, which summarizes the distribution of numeric attributes within a relatively narrow value range. The control menu provides a dynamic slider that can be used to restrict the selection to a subset of regions within a user-specified value range.

The *pie* chart (see Figure 3.34 for an example) is shown in addition to the area chart to summarize the distribution of numeric attributes that may span a wide value range. In

this case, clicking any segment of the pie chart opens a zoomed-in area chart specific for the genomic regions that fall into the selected value range.

The *neighborhood* chart (see Figure 3.30 for an example) illustrates the distribution of genome-wide maps — such as histone marks and transcription factor binding sites — in the vicinity of the selected region set. Average levels of overlap are calculated over all genomic regions in the set.

The *bubble* chart<sup>3</sup>(see Figure 3.20 for an example) plots the percentage of genomic regions that overlap with a given annotation (y-axis) against the total genome coverage of this type of annotation (x-axis). In this context, the genome coverage provides an indication of the expected overlap, highlighting annotations with substantially different overlap percentages. When used in comparison mode, we utilize the potential of a bubble chart to present any 3-dimensional data, by adding a dimension to represent the overlap of the annotations with the control set.

The *enrichment* chart (see Figure 3.33 for an example) summarizes gene-centric textual information in the form of a table and a word cloud. In the word cloud, the font size is scaled by the enrichment ratio, which is calculated relative to random expectation. Clicking on any annotation term refines the search to include only those regions that are associated with a gene carrying the corresponding annotation.

### Common actions when using the EpiExplorer user interface

*Uploading a custom dataset.* In the previous sections, we discussed how a custom dataset is preprocessed (see 3.3.2). In this section, we present the actions the user takes in order to upload a custom dataset into EpiExplorer. We assume the user has a correctly formatted BED file containing the list of genomic regions. Then the user needs to press the Upload button.

An upload page is shown where the user fills in various information about the dataset, such as dataset name, genome assembly, description (see Figure 3.14). He can additionally specify optional details such as if EpiExplorer should take strand information from the BED file into account and also what annotations EpiExplorer should use. By default, EpiExplorer uses all annotations. Finally, the user is asked if he wants to provide an email to use for notification. If the user chooses to provide an email, EpiExplorer automatically sends an email as soon as the preprocessing finishes. The email contains the dataset identifier and a direct link that loads the dataset into the EpiExplorer. If the user does not provide an email, once he finishes the upload he is provided a dialog window containing the URL location of a dynamic web page that notifies the user of the status of his dataset computation (see Figure 3.15). Before the dataset computation starts, EpiExplorer reports how many computations are in the queue ahead of the requested dataset. Once the dataset computation starts, the page is continuously updated with detailed information about the stage of the computation. After the preprocessing is complete, the page is updated to

---

<sup>3</sup>Implemented by Felipe Albrecht

Set custom annotation settings

Name of the dataset:
(\*)

Genome:
hg18

Dataset description:

Paste your dataset data here (in [BED format](#)):

(\*)

☐ Convert spaces to tabs

or select the dataset file:

(\*)

Use the strand data?
☐

Use the score data?
☐

Merge overlapping regions?
☐

Compute a reference?
☐

Email for automated notification:

Figure 3.14.: Uploading a custom dataset

contain the dataset identifier and a URL link that opens EpiExplorer with the custom dataset already loaded.

*Maintaining custom datasets.* EpiExplorer users commonly process more than one cus-

Your dataset is annotated by EpiExplorer at the moment!

**Step 1/6: Preparing the set of regions**

**Step 2/6: Preparing the annotations**

**Step 3/6: Computing annotation H3K4me2 (252 out of total 398 annotations are completed)**

Step 4/6: Exporting annotation data to text documents

Step 5/6: Building the CompleteSearch index

Step 6/6: Sending notification and cleaning up

Figure 3.15.: Dynamic status when preprocessing a custom dataset

tom dataset. These datasets can be loaded at different times and can be based on different genome assemblies. The EpiExplorer user interface helps the user by storing at his local computer the full list of his loaded datasets for each genome assembly. When the user opens EpiExplorer the datasets are automatically loaded in the backend and displayed in

the user interface. Similarly, when the user switches genome assemblies only the datasets corresponding to the selected genome assembly are loaded and all others are hidden.

*Sharing datasets and results.* Sharing and saving results from EpiExplorer is key to the usability of the software as was discussed in Section 3.2.1. EpiExplorer offers export options of two main types: dynamic and static. Dynamic exports aim to share either an EpiExplorer dataset or an exploration state that is saved with the purpose to be shared or explored later. The dynamic exports include dataset identifiers and analysis identifiers. The static exports have a different purpose, namely to facilitate integration with other tools. Static exports include exports of charts, region sets and gene-related lists.

The most commonly used export are dataset identifiers, which are discussed in multiple places in this chapter. The dataset identifiers enable any user who has the dataset identifier to load the dataset in EpiExplorer and generate any visualization, add and remove refinements and generally use all EpiExplorer features on it. To use a dataset identifier provided by a colleague, the user need to press the ‘Reload’ button and enter the identifier in the field that appears below.

Another type of dynamic export is saving a link to the current selection and visualization. This facilitates users to save their current state for documentation purposes and provides for sharing it with interested colleagues (as demonstrated by the supplementary tutorials of EpiExplorer (EpiExplorer: supplementary information, 2012)). To obtain a link to the current analysis, one needs to press the ‘Save as URL’ button below the main logo. A link appears under the top menu. This dynamically-generated URL can thereafter be used to reload the exact state of the current analysis.

The static exports include basic exporting and sharing of charts. Under most visualizations, the user can find the ‘To PNG’ button that immediately converts the visualization to a PNG image and offers the user to download it. Additionally, most charts have a ‘Show as table’ button below them that substitutes the visual chart with a table representing the chart data. This table is easily exported to spreadsheet software (i.e. Microsoft Excel or Google spreadsheets).

Other types of static exports refer to the current selection of regions. The easiest way to reach the full list of export options is to press on the ‘List the regions’ element immediately under the ‘Summary’ in the dataset view (see Figure 3.12). From there the user can export the current selection to a BED file, generate a URL that contains the listing of the regions, load it as custom track in the UCSC genome browser, load it with Ensembl, export it to Galaxy and to the Genomic HyperBrowser.

Finally, the user can export the gene-related properties of the current selection: nearby genes(within 5kb), the gene ontology terms and the OMIM terms associated with them. To do so, in the ‘Gene and annotations’ component of the dataset view (see Figure 3.12) the appropriate submenu must be selected. Next to the automatically generated table, there are export options that include: exporting gene identifiers, exporting Ensembl identifiers,

exporting gene symbols and exporting all GO terms associated with them.

*Comparing two datasets.* The comparison mode of EpiExplorer is used to explore the properties of two datasets together (see Section 3.2.1). In comparison mode, the current selection is the main dataset that is compared to a *reference* selection. There are several ways to choose a reference dataset. At any point, when exploring a selection the compare button can be chosen (see Figure 3.16). Immediately, the current dataset with all previ-

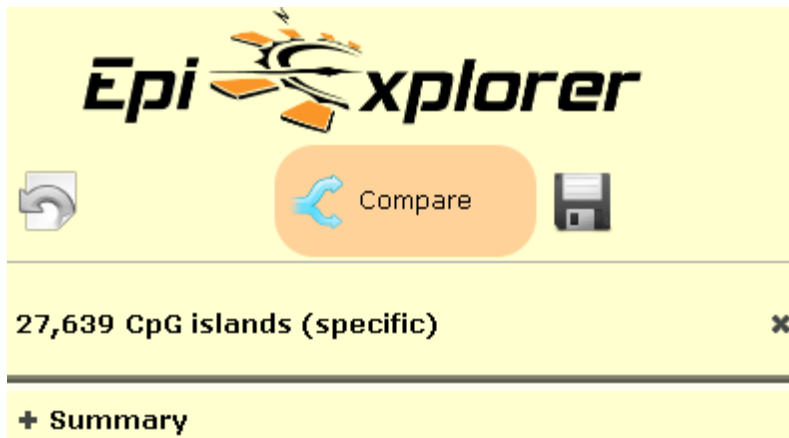


Figure 3.16.: The compare button (highlighted) activates the mode that allows direct comparison of two region selections

ous refinements is chosen as a reference dataset. Then, while the active dataset is being subjected to refinements, all visualizations will include also the reference dataset. Another way to select a reference dataset is to add it directly from the dataset selection screen (see Figure 3.11) by choosing the button to the left of the dataset name. The interface lists both the current selection as well as the reference dataset with all its refinements at all times (see Figure 3.17).

By default, a reference dataset is static. This means that once selected, the reference

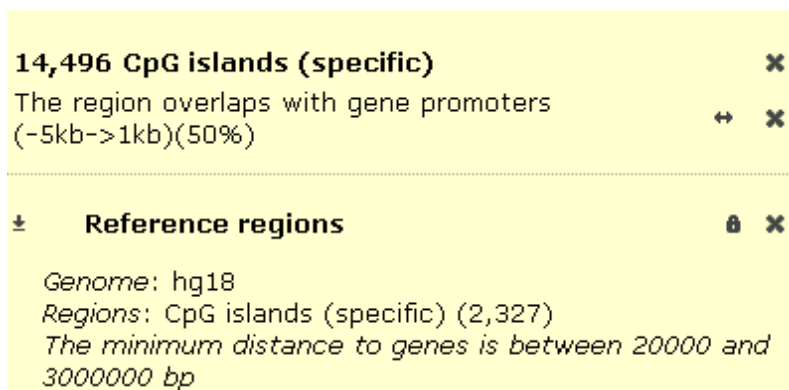


Figure 3.17.: At all times in comparison mode, the interface lists the current selection as well as the full details of the reference dataset

will remain unchanged, independently of any refinements that are added to the current

selection, removed from it or if the dataset of the current selection is changed altogether. The static reference is the default option. For more flexibility, a dynamic reference mode can be chosen. The dynamic mode synchronizes the active selection and the reference. Any refinement added to the current selection is also added to the reference selection. For example, the user wants to compare CpG islands located in gene promoters and compare their epigenetic properties to CpG islands located at least 20kb from the nearest gene (as in Figure 3.22). Imagine that as a next step, she wants to add to both region sets the refinement that the CpG islands do not overlap with any repeats and compare these two new region subsets. Using the dynamic reference mode (activated as in Figure 3.18), this is possible by simply adding the "no repeats" refinement to the active selection and it is automatically added to the reference. Also in dynamic mode, the user can remove any refinement from the reference (these are not removed from the current selection). The user may switch off the reference at any point by selecting the 'X' button on the right of the reference.

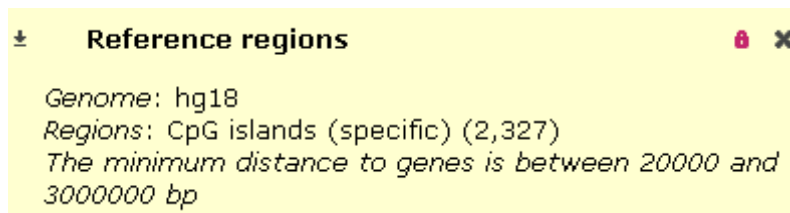


Figure 3.18.: Select the unlock button to activate the dynamic reference mode

## 3.4. Applications of EpiExplorer

In what follows, we use EpiExplorer in five different studies. First, we validate the software by presenting how one can rediscover known properties of CpG islands. Then, we explore the properties of 5hmC (5-hydroxymethylation), a novel epigenetic mark that may be important for CpG demethylation. We follow this up by narrowing down a large set of genomic locations to a strong candidate subset, useful for further experimental investigation. As a last biological study, we show that EpiExplorer can also be employed to analyze disease-specific data. More specifically, we compare the properties of locations of recurrent (consensus) breakpoints in several cancer tissue, to the properties of breakpoints, which were not observed in multiple samples. Finally, we present an overview of the activity on the EpiExplorer server during its first year as a public web service.

### 3.4.1. Rediscovering properties of CpG islands and application for discovery of robust CpG island annotations

In this section<sup>4</sup>, we report on a validation of the EpiExplorer method by studying the genome and epigenome characteristics of CpG islands, which is a relatively well-understood topic (Deaton and Bird, 2011). As outlined in the text below and the step-by-step online tutorial on the supplementary website (EpiExplorer: supplementary information, 2012), EpiExplorer makes it easy to rediscover the distinctive epigenetic characteristics of CpG islands, which have previously been studied using computational and experimental methods



(Bock et al., 2007; Cohen et al., 2011; Birney et al., 2007; Weber et al., 2007). The entire analysis can be performed in less than ten minutes without any bioinformatic training, guided by EpiExplorer’s context-specific visualizations. The exact steps leading to all results in this section are easy to verify using EpiExplorer, as described in a step-by-step tutorial on the Supplementary Website (EpiExplorer: supplementary information, 2012).

CpG islands account for some of the most important regulatory regions in the human genome (Deaton and Bird, 2011). These regions exhibit highly non-random epigenetic characteristics: on the one hand, most CpG islands are enriched for histone modifications indicative of open chromatin (e.g., H3K4me1 and H3K4me3), and they exhibit low levels of DNA methylation. On the other hand, specific subsets of CpG islands have been described as highly methylated or enriched for the repressive histone modification H3K27me3 (Bock et al., 2007; Cohen et al., 2011; Birney et al., 2007; Mendenhall et al., 2010; Strausman et al., 2009; Weber et al., 2007). In order to validate EpiExplorer on the well-studied topic of epigenetic regulation at CpG islands, here we analyze the characteristics of CpG islands across the human genome, using EpiExplorer’s functionality for exploring genomic region sets in the context of public genome and epigenome datasets.

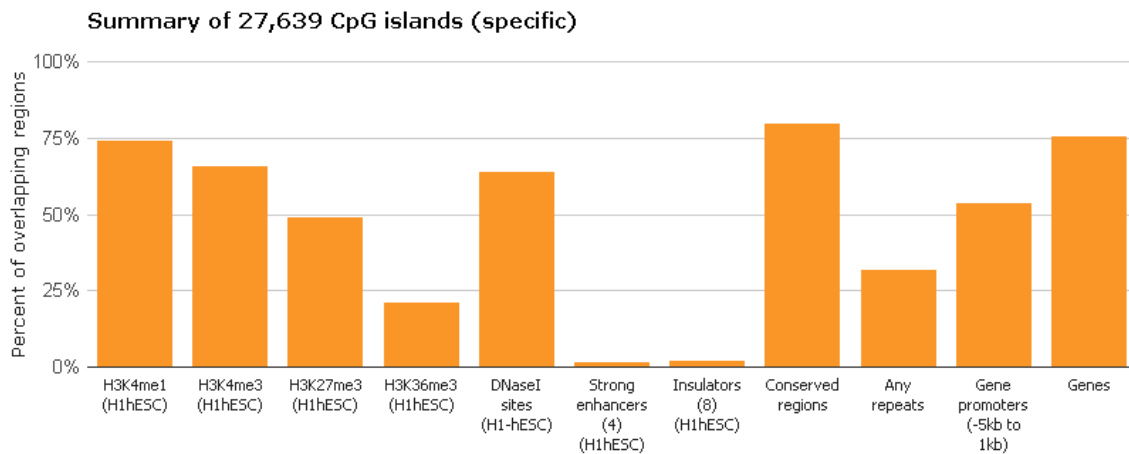


Figure 3.19.: Bar chart summarizing the percent overlap (y-axis) between CpG islands and various genomic region sets (x-axis) in H1hESC cells.

CpG islands are already available as one of EpiExplorer’s default region sets, hence it is not necessary to upload any new dataset to perform this analysis. Once we select ‘CpG islands (specific)’ from the exploration menu on the left of EpiExplorer’s start screen, EpiExplorer displays a summary of genome and epigenome annotations that co-localize with CpG islands (Figure 3.19). According to this diagram, more than half of all CpG islands overlap with Ensembl-annotated gene promoter regions, which is a well-established observation in the literature (Bajic et al., 2006; Ioshikhes and Zhang, 2000). Furthermore, we observe that two-thirds of all CpG islands overlap with the promoter-associated histone H3K4me3 mark in ES cells (Figure 3.19) and in other tissues (Figure 3.20). This observation underlines that a large subset of CpG islands indeed carry the key chromatin mark indicative of active promoters, and it constitutes a substantial enrichment as genomic

<sup>4</sup>The work presented in this section was published as part of Halachev et al. (2012).

regions carrying this mark cover only two to four percent of the genome (Figure 3.20). Furthermore, EpiExplorer’s neighborhood plot (Figure 3.21) highlights how strongly and specifically the H3K4me3 mark is enriched at the boundaries of CpG islands compared to the broader genomic neighborhood of these regions.

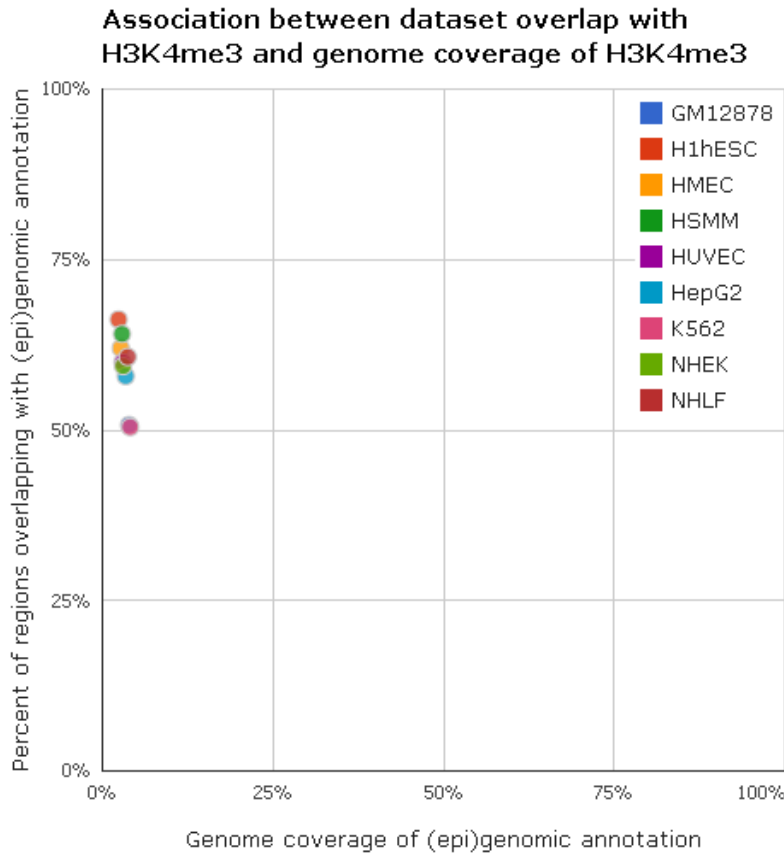


Figure 3.20.: Bubble chart plotting the percent overlap (y-axis) between CpG islands and H3K4me3 peaks in specific tissues (color-coded) against the total genomic coverage of all corresponding peaks (x-axis)

While association with open chromatin appears to be the default state of most bona fide CpG islands in the human genome (Bock et al., 2007; Cohen et al., 2011; Birney et al., 2007; Mendenhall et al., 2010; Straussman et al., 2009), it has been shown that a subset of CpG islands are frequently associated with the repressive histone H3K27me3 mark (Ku et al., 2008; Mikkelsen et al., 2007). CpG islands have also even been reported to play a role in recruiting Polycomb proteins and the H3K27me3 mark in ES cells (Mendenhall et al., 2010). An EpiExplorer neighborhood plot shows specific and localized enrichment of the H3K27me3 mark in a human ES cell line, with an enrichment peak that ranges from one kilobase upstream of the annotated CpG island borders to one kilobase downstream (Figure 3.22). This ES-cell specific enrichment peak is only marginally broader than the one observed for the H3K4me3 mark (Figure 3.21). In contrast, for cell types other than ES cells, we observe elevated levels of H3K27me3 in a broad neighborhood surrounding CpG islands, consistent with the observation that localized peaks of H3K27me3 in ES cells are resolved into broad H3K27me3-enriched broad local enrichments in differentiated cells

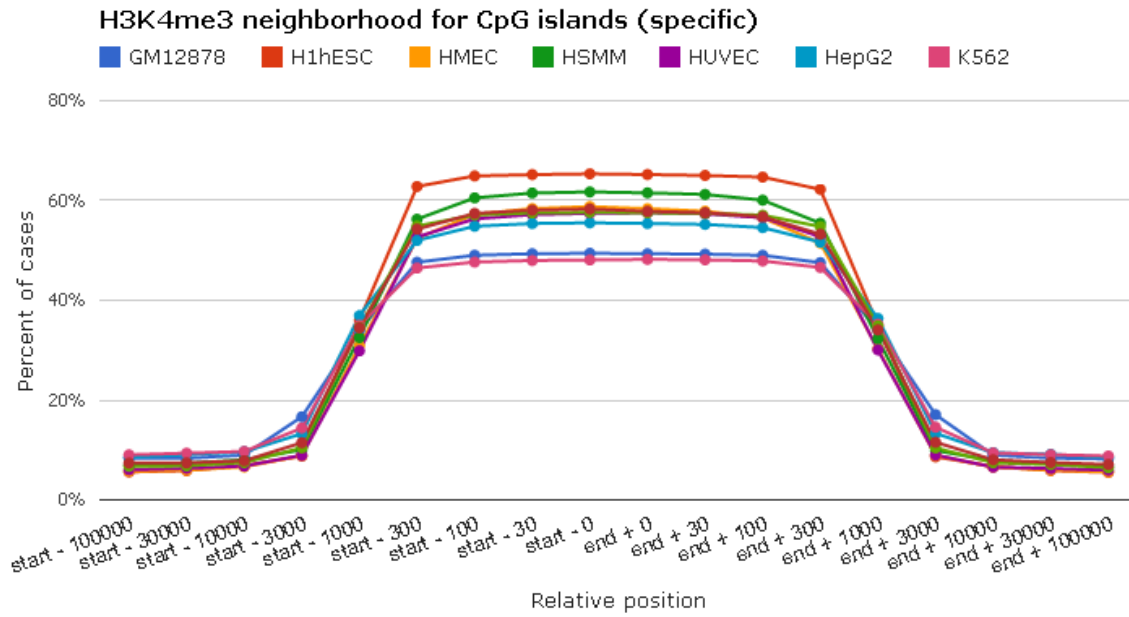


Figure 3.21.: Neighborhood plot illustrating the percent overlap (y-axis) with histone H3K4me3 peaks in the vicinity of CpG islands (x-axis). Line colors correspond to histone modification data for different cell types.

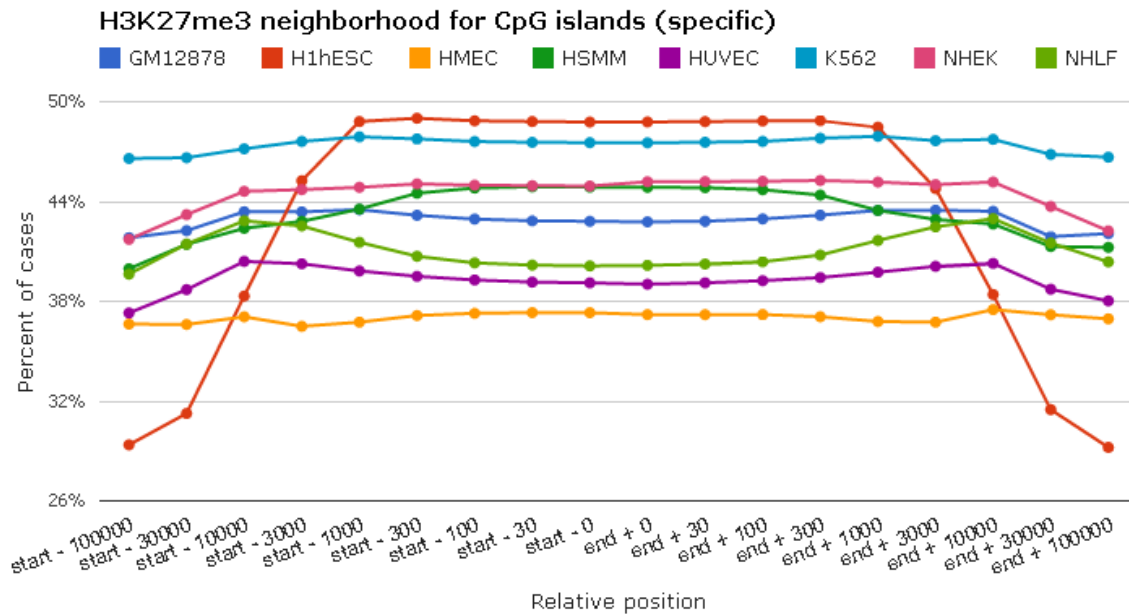


Figure 3.22.: Neighborhood plot illustrating the percent overlap (y-axis) with histone H3K27me3 peaks in the vicinity of CpG islands (x-axis). Line colors correspond to histone modification data for different cell types.

(Pauler et al., 2009).

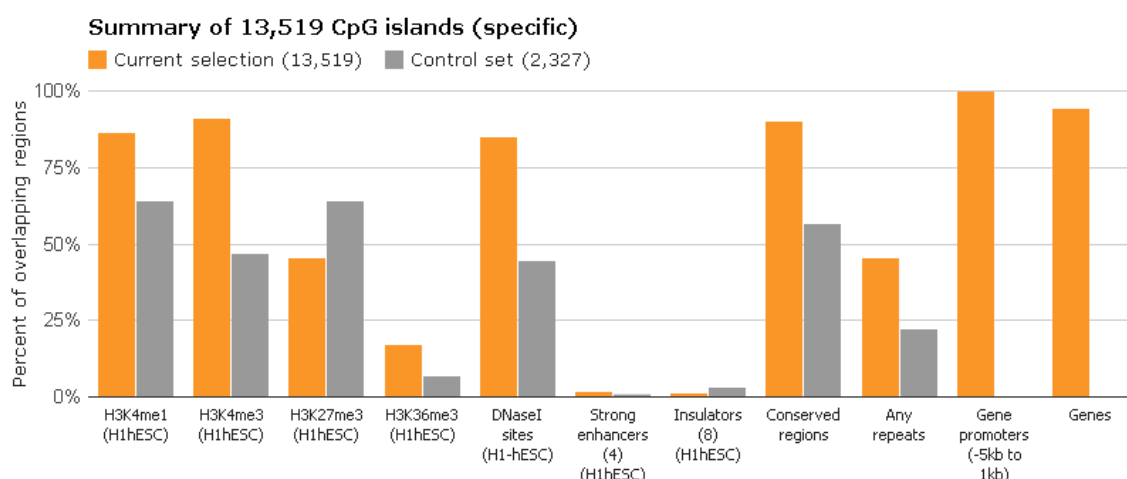


Figure 3.23.: Percent overlap (y-axis) of 13,519 CpG islands located within one kilobase from a gene transcription start site (orange) and 2,327 CpG islands located at least 20 kilobases from the nearest gene (grey) with genome and epigenome annotation data (x-axis)

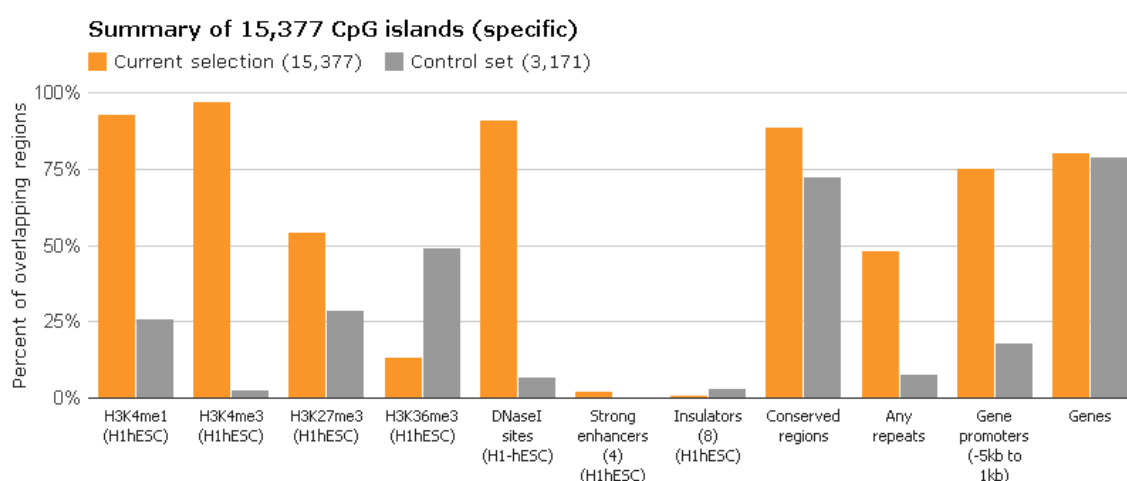


Figure 3.24.: Percent overlap (y-axis) of 15,377 constitutively unmethylated CpG islands (orange, less than 30% methylation in seven tissues) and 3,171 constitutively methylated CpG islands (grey, more than 60% methylation in the same seven tissues) with genome and epigenome annotation data (x-axis).

Despite the strong overlap of CpG islands with gene promoters and other genic regions (Figure 3.19), almost a quarter of CpG islands (6,705 in total) do not overlap with any annotated promoter regions or genes and are therefore categorized as intergenic CpG islands (Illingworth et al., 2010). Some of these intergenic CpG islands may be linked to genes and promoters that are currently missed by genome annotations (e.g. lincRNAs), but others may have non-canonical roles for example as distal enhancers or as anchor points in the maintenance of three-dimensional genome organization. Using the refinement tools of EpiExplorer, we can dynamically reduce the set of all CpG islands to those that are

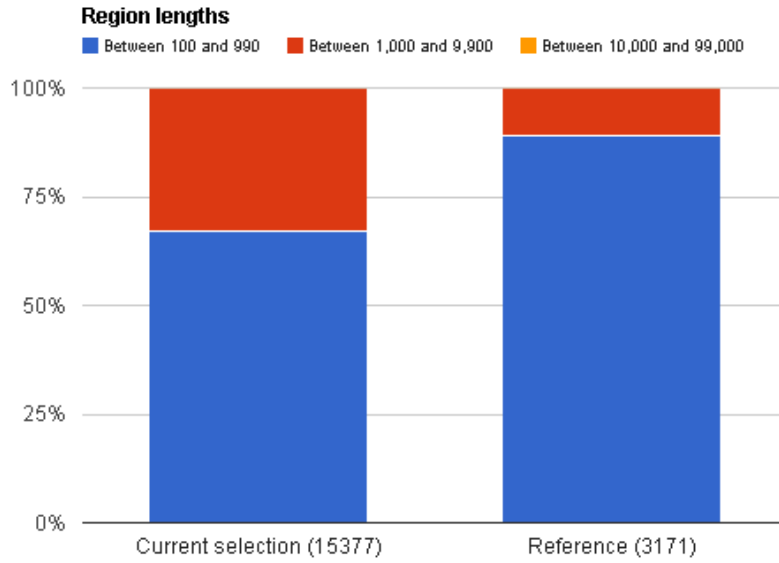


Figure 3.25.: Overview of the length distribution of constitutively unmethylated CpG islands (left) and constitutively methylated CpG islands (right).

clearly intergenic (i.e. located at least 20 kilobases distant from the nearest gene) and compare their properties with those of a set of promoter-associated CpG islands (i.e. located within a kilobase of an annotated transcription start site). The results show that intergenic CpG islands less frequently exhibit the promoter-associated H3K4me3 mark and the transcription-associated H3K36me3 mark than promoter-associated CpG islands (Figure 3.23), consistent with their intergenic nature. On the other hand, they are associated more frequently with H3K27me3 peaks and insulator elements (Figure 3.23), which is suggestive of a structural role in the organization of chromatin. We also explored the distribution of DNA methylation among CpG islands. While most CpG islands appear to be unmethylated in the germline and thus protected from the increased C-to-T mutation rates associated with cytosine methylation (Cohen et al., 2011; Bock et al., 2006; Smallwood et al., 2011), a subset of CpG islands becomes methylated during somatic tissue differentiation (Meissner et al., 2008; Mohn et al., 2008). Furthermore, certain types of repeat-associated and exonic CpG islands appear to be methylated in all tissues and retain their moderate levels of CpG density by means other than the absence of DNA methylation in the germline (Cohen et al., 2011; Maunakea et al., 2010). To compare the genomic characteristics of methylated and unmethylated CpG islands, we derived within EpiExplorer a test set of constitutively unmethylated CpG islands and a reference set of constitutively methylated CpG islands (Figure 3.24). Comparison of both types (unmethylated CpG islands shown in orange, methylated ones in gray) identified striking enrichment for open-chromatin associated marks (H3K4me1, H3K4me3, DNaseI hypersensitive sites) among unmethylated CpG islands. In contrast, methylated CpG islands were strongly associated with the transcription-linked H3K36me3 mark and exhibited a similar level of evolution-

ary conservation and gene association as unmethylated CpG islands. The characteristic

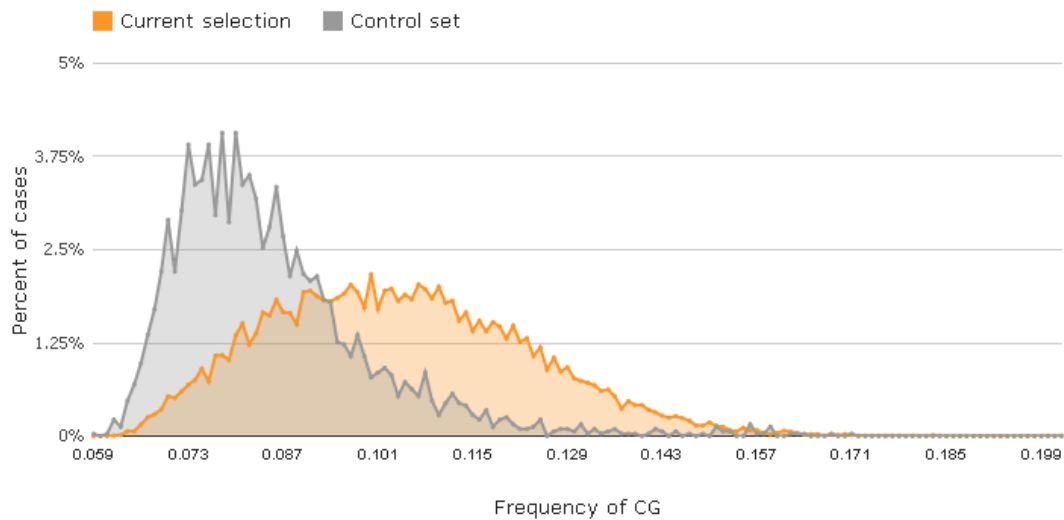


Figure 3.26.: Distribution of CpG dinucleotide frequencies among constitutively unmethylated CpG islands (orange) and among constitutively methylated CpG islands (grey).

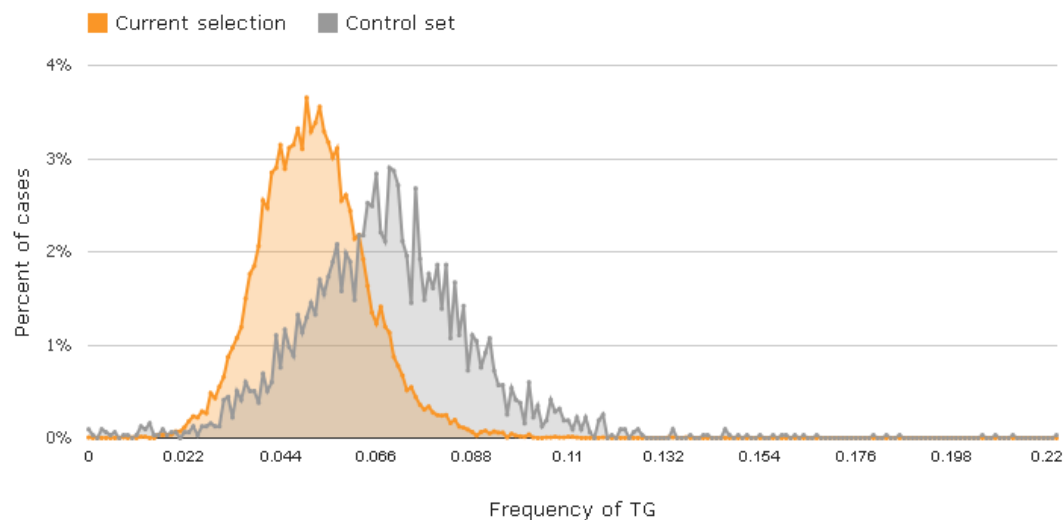


Figure 3.27.: Distribution of TpG dinucleotide frequencies among constitutively unmethylated CpG islands (orange) and among constitutively methylated CpG islands (grey).

differences between unmethylated and methylated CpG islands are not limited to their genomic location relative to genes and chromatin marks, but also include the genomic DNA sequence of the CpG islands themselves. Consistent with previous reports that identified

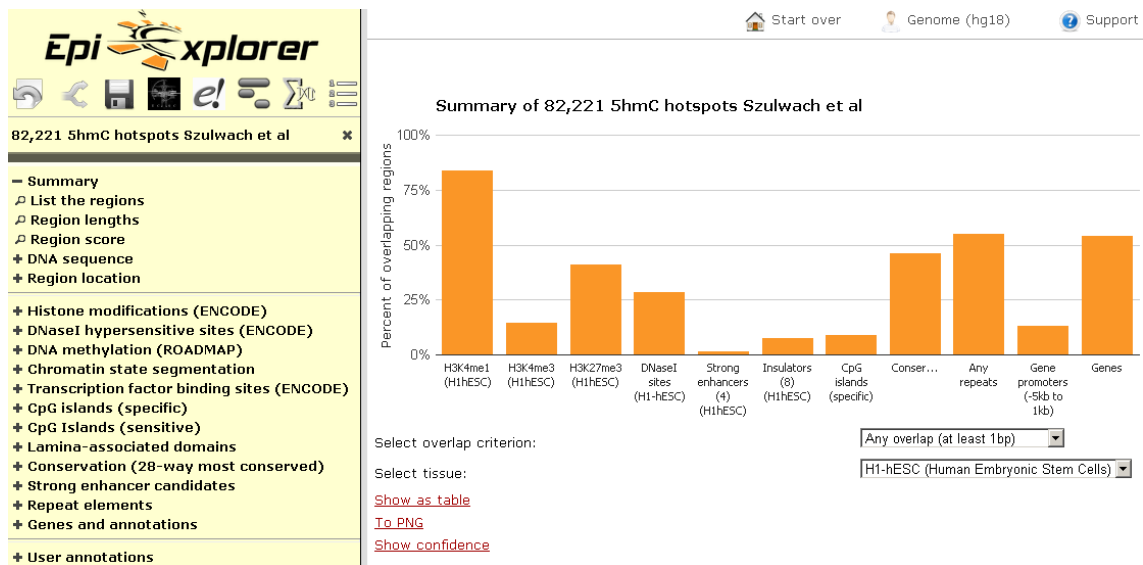


Figure 3.28.: Bar chart summarizing the percent overlap (y-axis) between 5hmC hotspots and various genomic datasets (x-axis) in H1hESC cells.

high CpG island length and CpG density as strong predictors of low DNA methylation levels (Straussman et al., 2009; Weber et al., 2007; Bock et al., 2006; Das et al., 2006), the EpiExplorer analysis shows that unmethylated CpG islands tend to be longer (Figure 3.25) and exhibit a CpG density distribution that is substantially shifted toward increased CpG densities compared to their methylated counterparts (Figure 3.26). In contrast, the TpG density distribution shows an opposite trend (Figure 3.27), supporting the notion that high levels of DNA methylation are directly linked to the accumulation of C-to-T mutations in the germline. In summary, these observations suggest that CpG islands are regulated in different ways by three epigenetic marks, histone H3K4me3, histone H3K27me3 and DNA methylation. The presence of H3K4me3 is strongly correlated with low levels of DNA methylation. Furthermore, H3K27me3 overlaps with H3K4me3 at a subset of CpG islands (in particular for ES cells), while co-localization between H3K27me3 and DNA methylation is rare and only observed at CpG islands that are not particularly CpG-rich.

### 3.4.2. Connecting a new epigenetic mark to reference maps of the human genome and epigenome

#### Discovery of properties of 5hmC

To assess the utility of EpiExplorer for exploratory analysis and hypothesis generation in a more advanced setting, we investigated a recently discovered epigenetic mark<sup>5</sup>. 5-Hydroxymethylcytosine (5hmC) is a chemical variant of normal (that is, non-hydroxylated) cytosine methylation. It was first observed in embryonic stem (ES) cells and in certain types of neurons (Kriaucionis and Heintz, 2009; Tahiliani et al., 2009). The conversion of cytosine methylation into 5hmC is catalyzed by proteins of the TET family. One TET protein (TET2) is frequently mutated in myeloid cancers (Delhommeau et al., 2009), underlining

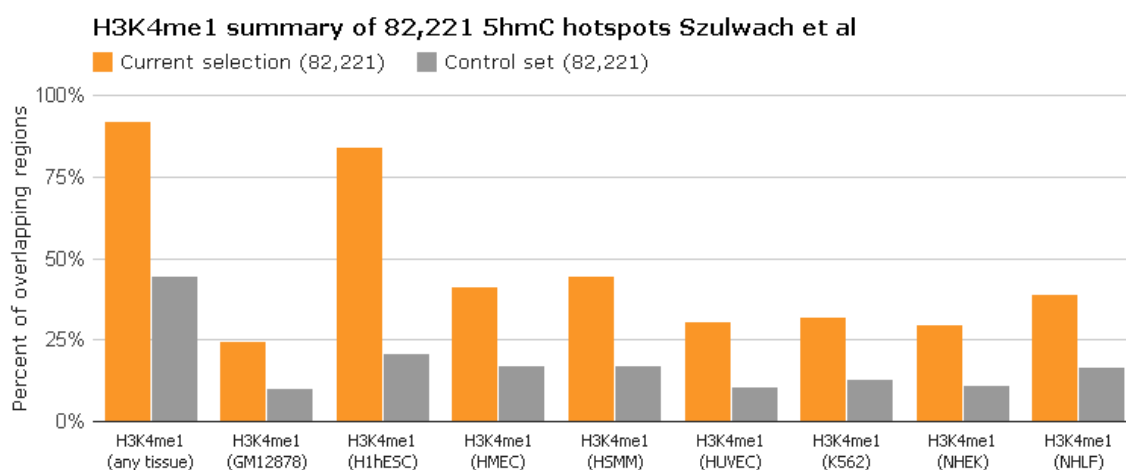


Figure 3.29.: Bar chart comparing the percent overlap of 5hmC hotspots (orange) and randomized control regions (grey) with histone H3K4me1 peaks, based on ENCODE data (Myers et al., 2011).

the biomedical relevance of studying the role of 5hmC in gene regulation.

From the paper of Szulwach et al. (2011), we obtained the genomic region coordinates for a total of 82,221 hotspots of 5hmC that the authors experimentally mapped in human ES cells. We uploaded these hotspot regions into EpiExplorer, where they are automatically annotated with default genomic attributes such as gene annotations and associated epigenetic marks. EpiExplorer’s initial overview screen summarizes the overlap of 5hmC hotspots with the most relevant genomic attributes and provides the starting point for interactive exploration of the dataset (Figure 3.28). This view is tissue-specific, and we select a human ES cell line (‘H1hESC’) as the tissue type of interest. In ES cells, we observe striking overlap between 5hmC hotspots and epigenetic marks associated with distal gene-regulatory activity. Specifically, more than 80% of the 5hmC hotspots overlap with peaks of the histone H3K4me1 mark, which is a well-known signature of enhancer elements (Heintzman et al., 2009). In contrast, less than 20% of 5hmC hotspots overlap with histone H3K4me3 (Figure 3.28), which is considered the hallmark of active core promoter regions (Kouzarides, 2007).

To assess whether the association of 5hmC hotspots with H3K4me1 peaks indeed constitutes a relevant enrichment, we performed the same comparison for a randomized control set. EpiExplorer automatically calculates such control sets for user-uploaded region sets, which is done by reshuffling the genomic positions while retaining the overall number of regions and the distribution of region sizes. Visual comparison shows that the overlap between 5hmC hotspots and H3K4me1 peaks is indeed fourfold higher than expected by chance (Figure 3.29), constituting a strong enrichment with potential biological implications. This enrichment is much more pronounced for H3K4me1 in ES cells than for other tissues, supporting the specificity of the observed association. We could further validate this association using EpiExplorer’s neighborhood plot (Figure 3.30). When plotting the

<sup>5</sup>The work presented in this section was published as part of Halachev et al. (2012).



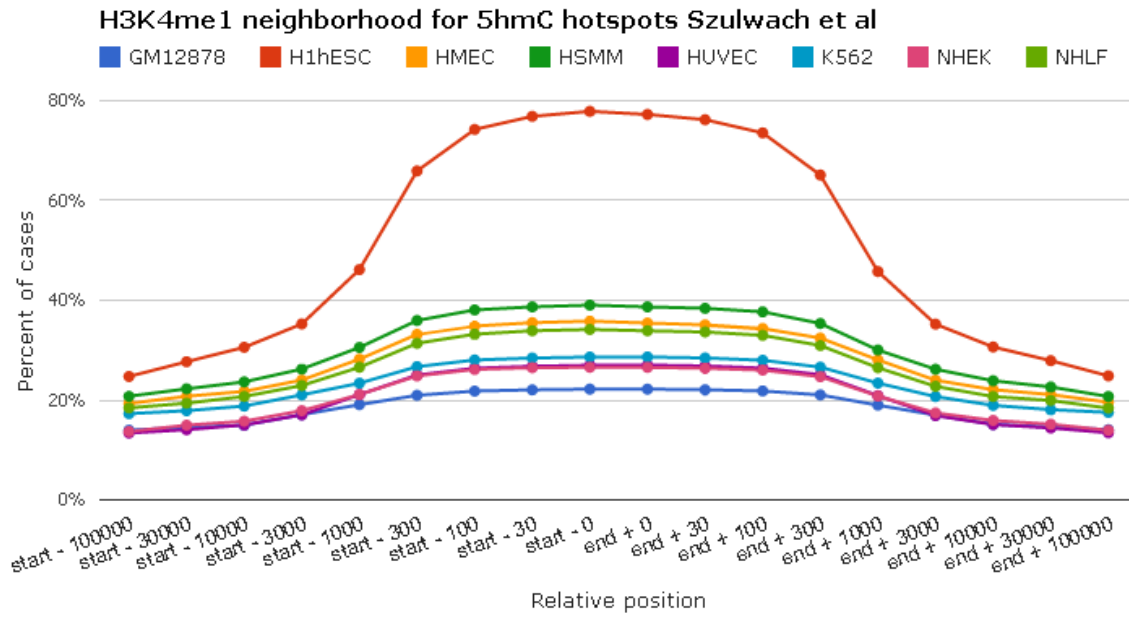


Figure 3.30.: Genomic neighborhood plot illustrating the percent overlap (y-axis) with H3K4me1 peaks in the vicinity of 5hmC hotspots (x-axis). Different line colors correspond to H3K4me1 data for different cell types.

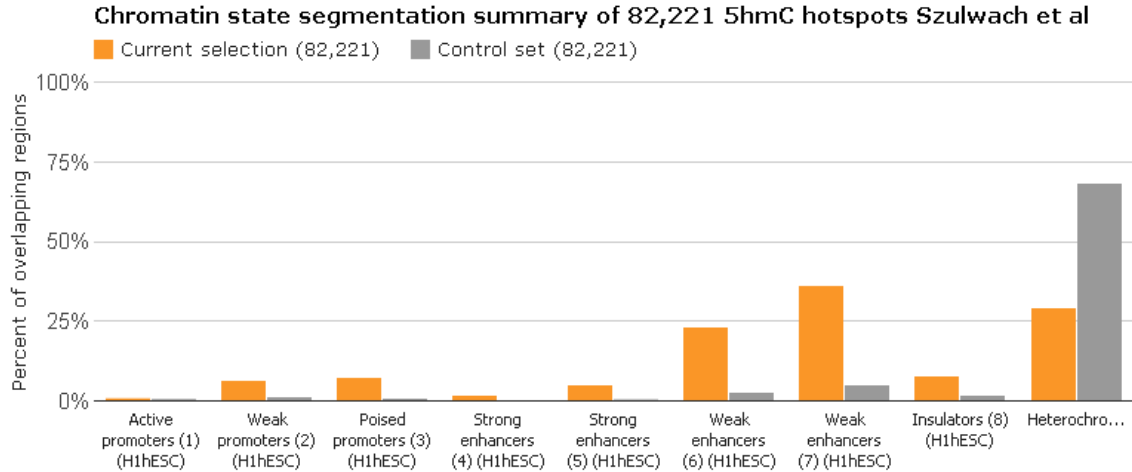


Figure 3.31.: Bar chart comparing the percent overlap of 5hmC hotspots (orange) and randomized control regions (grey) with a comprehensive catalog of epigenetic states derived by computational segmentation of ENCODE histone modification data (Ernst et al., 2011).

levels of H3K4me1 methylation in the vicinity of 5hmC hotspots across the genome, we again observed a much stronger enrichment for ES cells than for H3K4me1 data from other tissues. Furthermore, when we compared the 5hmC hotspots with a comprehensive catalog of epigenetic states (Ernst et al., 2011), we detected striking enrichment for several classes

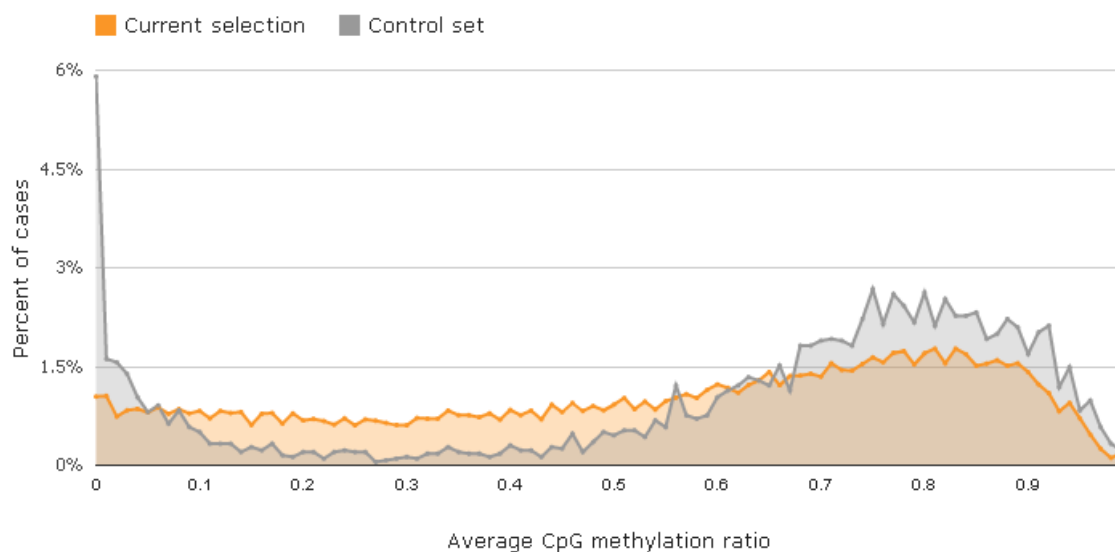


Figure 3.32.: Distribution of DNA methylation levels among 5hmC hotspots (orange) and randomized control regions (grey), based on Roadmap Epigenomics data (Human Epigenome Atlas, 2013).

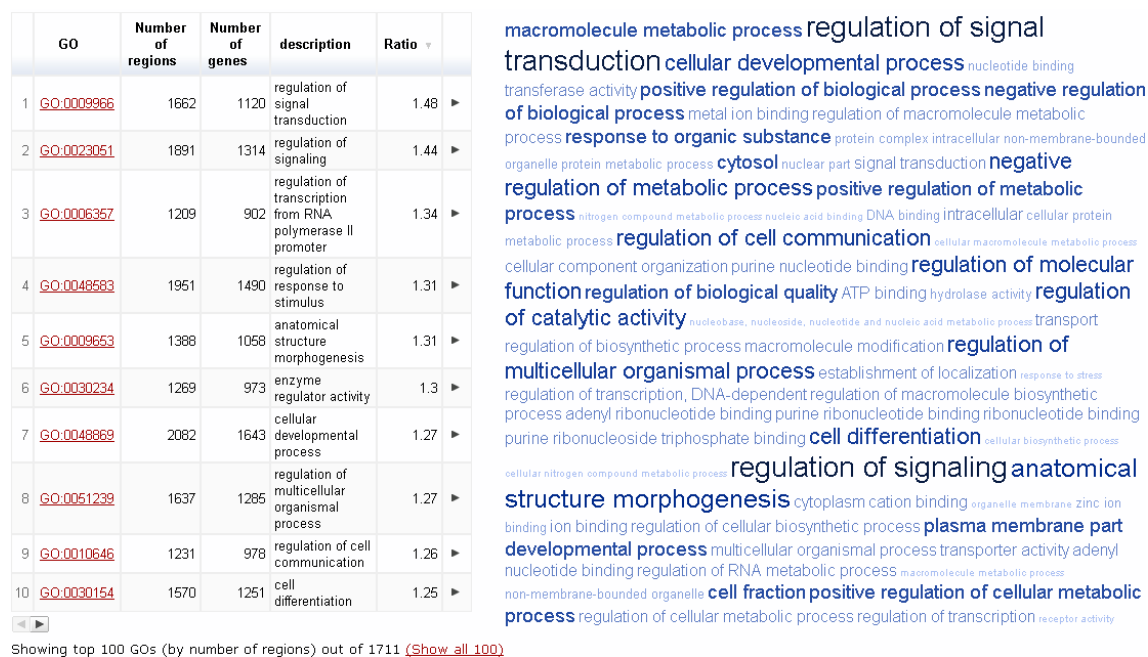


Figure 3.33.: Enrichment table (left) and word cloud (right) illustrating the most highly enriched Gene Ontology (GO) terms among genes whose transcribed region is within 10 kb of a 5hmC hotspot. The most general (more than 5,000 associated genes) and most specific GO terms (less than 50 associated genes) were suppressed in this analysis.

of enhancer elements (Figure 3.31). In summary, these results suggest the hypothesis that a specific association may exist between 5hmC and H3K4me1-marked enhancer elements in human ES cells.

Given the presumed role of 5hmC in the erasure of DNA methylation (Mohr et al., 2011; Munzel et al., 2011), we also investigated the distribution of normal (that is, non-hydroxylated) cytosine methylation among 5hmC hotspots, again in comparison with the randomized control set. To that end, we use the ability of EpiExplorer to work on dynamically refined subsets of the data and filter the set of 5hmC hotspots down to those regions for which we also have sufficient DNA methylation data (a step-by-step tutorial is available at EpiExplorer: supplementary information (2012)). The results show that 5hmC hotspots are rarely unmethylated but frequently associated with moderate levels of DNA methylation in the range of 10% to 50% (Figure 3.32), which is consistent with significant but incomplete demethylation activity occurring at the majority of 5hmC hotspots. This observation is also supported by a recent report describing enrichment of 5hmC and enhancer activity in genomic regions with intermediate DNA methylation (Stadler et al., 2011). Finally, we use EpiExplorer to perform a Gene Ontology analysis for those genes that are located in close vicinity of 5hmC hotspots (Figure 3.33). The 5hmC-associated genes are enriched for specific annotation terms related to gene regulation and development, including ‘regulation of signal transduction’, ‘cell differentiation’ and ‘anatomical structure morphogenesis’.

Taken together, these EpiExplorer analyses suggest testable hypotheses about the role of 5hmC in human ES cells. For example, active DNA demethylation – with 5hmC as an intermediate – may protect developmental enhancers from gaining DNA methylation in undifferentiated cells. This mechanism may help ES cells retain their developmental potential in the presence of high levels of DNA methyltransferase activity. In addition, active DNA methylation could help avoid the accumulation of cancer-associated epigenetic alterations in undifferentiated cells, given that the sites of such alterations frequently overlap with developmental regulatory elements (De Carvalho et al., 2010). To provide further support for these hypotheses, we can export the analyzed data from EpiExplorer to the Genomic HyperBrowser and perform more rigorous statistical testing than is possible within EpiExplorer. And most importantly, it will be necessary to confirm biological significance by in-depth functional dissection of the interplay between 5hmC and H3K4me1 at developmental enhancers. Such wet-lab studies are laborious to conduct and inherently limited to a small number of candidate genes or genomic regions, thus requiring careful selection of the most relevant candidates. EpiExplorer can help guide the selection of suitable regions for functional follow-up, as illustrated in the following case study.

### **Interactive identification and prioritization of candidate regions using EpiExplorer**

When studying mechanisms of gene regulation, it is often necessary to select a few model genes or genomic regions in order to afford a more detailed investigation than is possible with genome-wide methods. Good candidates should be informative of the phenotype of interest but must also be easily tractable experimentally. EpiExplorer is a powerful tool for identifying such candidates through several steps of region set filtering and interactive refinement of the selection criteria. For example, to unravel the mechanistic basis of the association between 5hmC and H3K4me1-marked enhancer elements (as described in the previous section) we need to identify a handful of strong examples for this kind of association, which can then be studied using biochemical and molecular biological assays.

Good candidate regions should exhibit robust enrichment for both 5hmC and H3K4me1, proximity to genes involved in transcriptional regulation, and moderate levels of DNA methylation. With EpiExplorer, it is straightforward to distill such candidate regions from the complete list of 82,221 5hmC hotspots.

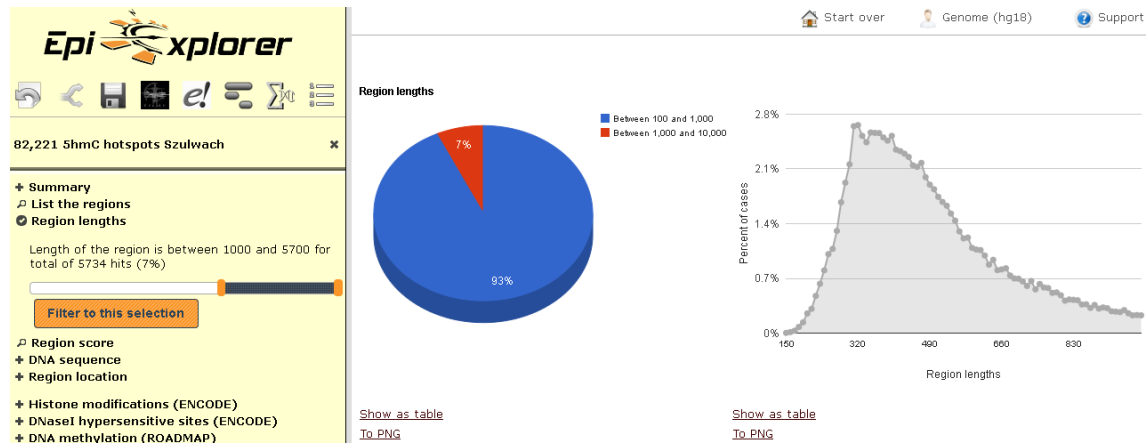


Figure 3.34.: Using successive filtering steps, a genomic dataset with 82,221 hotspots of 5-hydroxymethylcytosine (5hmC) in human ES cells (Szulwach et al., 2011) is refined to a list of 16 regions that provide strong candidates for investigating the functional association between 5hmC and H3K4me1-marked enhancer elements. (a) Filtering with a minimum length threshold of 1 kb yields 5,734 genomic regions.

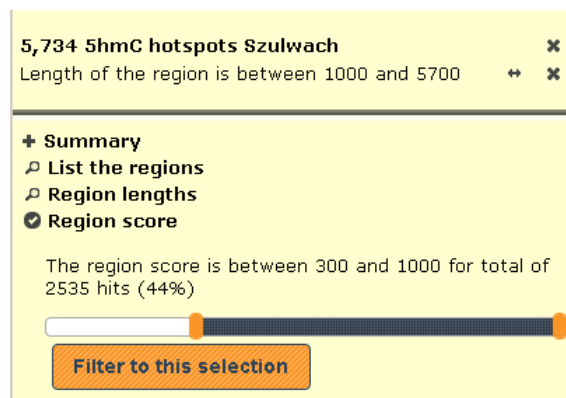


Figure 3.35.: Filtering with a minimum 5hmC hotspot score threshold of 300, which corresponds to a detection significance of 10-30 or better, yields 2,535 genomic regions.

First, we inspect the length distribution of 5hmC hotspots (Figure 3.34) and retain only those hotspots with a minimum length of 1 kb, which removes spurious peaks that are occasionally introduced by short repetitive elements in the genomic DNA sequence. Second, we filter for a detection significance of 10-30 or better in order to focus the analysis on the most clear-cut 5hmC hotspots (Figure 3.35). Third, we require evidence of an enhancer-associated chromatin signature and retain only those 5hmC hotspots that overlap with

- Histone modifications (ENCODE)		
- H3K4me1		
Neighborhood		
<input checked="" type="checkbox"/> Overlapping		
H3K4me1 (GM12878)	597	
H3K4me1 (H1hESC)	2,334	
H3K4me1 (HMEC)	1,173	
H3K4me1 (HSMM)	1,134	
H3K4me1 (HUVEC)	850	
H3K4me1 (K562)	810	
H3K4me1 (NHEK)	768	
H3K4me1 (NHLF)	1,004	
H3K4me1 (any tissue)	2,418	
<input type="checkbox"/> Not overlapping		
+ Distance to nearest		

Figure 3.36.: Filtering for overlap with H3K4me1 peaks in a human ES cell line (H1hESC) yields 2,334 genomic regions.

+ Chromatin state segmentation		
+ CpG islands (specific)		
+ CpG islands (sensitive)		
+ Conservation		
+ Repeat elements		
- Genes and annotations		
<input type="checkbox"/> Gene names (Ensembl)		
<input type="checkbox"/> Gene ontology (terms)		
<input checked="" type="checkbox"/> Gene ontology (words)		
<input type="checkbox"/> OMIM (terms)		
<input type="checkbox"/> OMIM (words)		
+ Genes		
+ Gene promoters		
+ Gene transcription start sites		
+ Gene exons		
+ User annotations		

Word	Number of GOs with such description	
1 regulation	1608	
2 activity	1066	
3 process	835	
4 cell	520	
5 positive	482	
6 binding	423	
7 negative	389	
8 metabolic	373	
9 protein	362	
10 response	343	

Figure 3.37.: Filtering for association with genes that are annotated with any of the 1,608 Gene Ontology terms containing the word 'regulation' yields 1,064 genomic regions.

H3K4me1 peaks (Figure 3.36). Fourth, in order to maximize relevance of the candidate regions for drawing conclusions about gene regulation, we restrict the analysis to genomic regions located in the vicinity of genes that are annotated with Gene Ontology terms containing the word 'regulation' (Figure 3.37). Fifth, we import an additional dataset of 5hmC hotspots in human ES cells (Stroud et al., 2011) into EpiExplorer and retain only those hotspots that are present in both datasets (Figure 3.38). Because these two 5hmC datasets were obtained using different experimental methods, our selection of consensus hotspots should effectively remove technical artifacts of either dataset. Sixth, to be able to robustly select 5hmC hotspots with intermediate DNA methylation levels in the last step, we discard those regions for which insufficient bisulfite sequencing coverage is available from the Roadmap Epigenomics datasets (Figure 3.39). Seventh and last, we focus the analysis on those regions that exhibit moderate levels of DNA methylation because it is plausible to hypothesize that the epigenetic state of these regions might be the result of significant but incomplete levels of active DNA demethylation (Figure 3.40). Each of these filtering steps

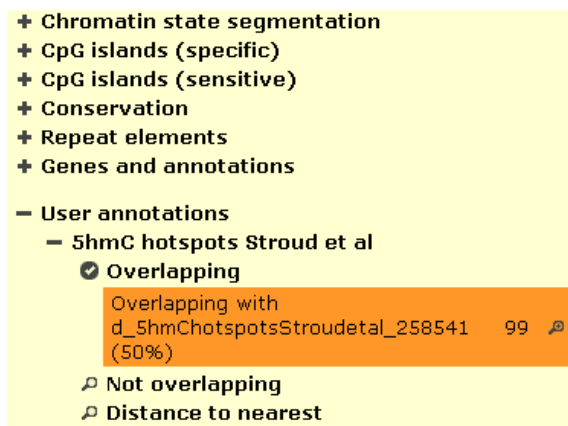


Figure 3.38.: Filtering for overlap with an alternative dataset of 5hmC hotspots (Stroud et al., 2011) yields 99 genomic regions.

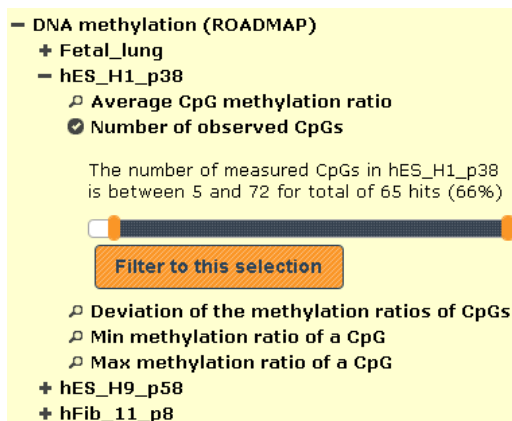


Figure 3.39.: Filtering for a minimum DNA methylation coverage threshold of five CpGs yields 65 genomic regions.

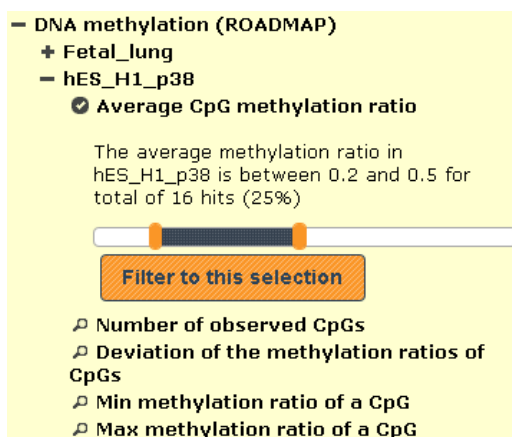
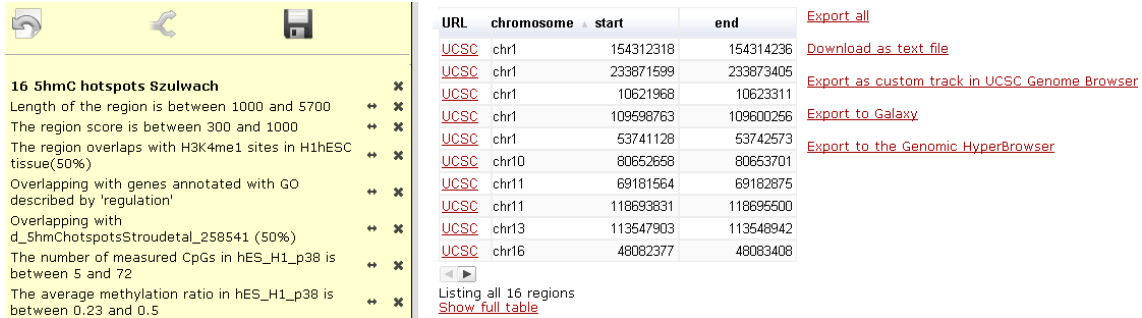


Figure 3.40.: Filtering for intermediate DNA methylation with levels in the range of 20% to 50% yields 16 genomic regions.

is interactively performed using EpiExplorer. Together they bring the original list of 82,221 5hmC hotspots down to 16 regions that fulfill all criteria and constitute strong candidates for a mechanistic study exploring the association between 5hmC and H3K4me1-marked enhancer elements (Figure 3.41).



The screenshot shows the EpiExplorer interface. On the left, a sidebar lists 16 5hmC hotspots with various filters and a 'Show full table' link. The main panel displays a table of genomic regions with columns for URL, chromosome, start, and end. To the right of the table are links for exporting the data to various tools.

URL	chromosome	start	end
<a href="#">UCSC</a>	chr1	154312318	154314236
<a href="#">UCSC</a>	chr1	233871599	233873405
<a href="#">UCSC</a>	chr1	10621968	10623311
<a href="#">UCSC</a>	chr1	109598763	109600256
<a href="#">UCSC</a>	chr1	53741128	53742573
<a href="#">UCSC</a>	chr10	80652658	80653701
<a href="#">UCSC</a>	chr11	69181564	69182675
<a href="#">UCSC</a>	chr11	118693831	118695500
<a href="#">UCSC</a>	chr13	113547903	113548942
<a href="#">UCSC</a>	chr16	48082377	48083408

Export all  
Download as text file  
Export as custom track in UCSC Genome Browser  
Export to Galaxy  
Export to the Genomic HyperBrowser

Listing all 16 regions  
Show full table

Figure 3.41.: EpiExplorer screenshot showing the final list of candidate regions, ready for visualization in a genome browser, for download and manual inspection, and for export to other web-based tools for further analysis.

To facilitate follow-up research, EpiExplorer provides extensive functionality for data export and visualization using external tools. First, every genomic region set in EpiExplorer can be exported and visualized as a custom track in the UCSC Genome Browser (Karolchik et al., 2008), which is usually a good starting point for designing locus-specific experiments. Second, the results generated by EpiExplorer can be transferred to Galaxy (Goecks et al., 2010) in order to perform sequence motif search, primer design and a number of other useful analyses that facilitate wet-lab experimental planning. Third, export to EpiGRAPH (Bock et al., 2009) or the Genomic HyperBrowser (Sandve et al., 2010) can provide the starting point for additional statistical analyses performed online. Fourth, it is possible to export and download all region sets as text files for customized analysis with spreadsheet software (for example, Excel) or statistical analysis tools (for example, R).

### 3.4.3. Epigenetic properties of cancer breakpoints

In this section<sup>6</sup>, we present results described in Tološi et al. (2013). In that study, our colleagues define consensus breakpoints in cancer as genomic locations around which copy number breakpoints occur more frequently than expected by chance. A method is proposed (called C-KS) for identification of consensus breakpoints. The method is applied to several cancer arrayCGH datasets: breast, colon, ovarian, neuroblastoma and glioblastoma. We used EpiExplorer for a qualitative validation of the C-KS algorithm: we investigate genomic and epigenomic properties of consensus breakpoints and show that they tend to be enriched in functional elements and certain DNA sequence patterns. More importantly, we demonstrate the usability of EpiExplorer in disease-specific studies.

<sup>6</sup>The work presented in this section was published as part of Tološi et al. (2013).

Dataset	No. of samples	No. of breakpoints per sample	No. of consensus breakpoints identified by C-KS
Neuroblastoma	162	54	62
Colon	98	168	173
Glioblastoma	539	261	492
Breast173	173	339	320
Breast54	54	394	327
Breast167	167	461	503
Ovarian	290	806	662

Table 3.1.: Number of tumor, samples, breakpoints and consensus breakpoints identified in sever cancer datasets.

### Algorithm for identification of consensus breakpoints and application to cancer data

Tološi et al. (2013) introduces the Consensus breakpoints by Kernel Smoothing (C-KS) algorithm for identifying recurrent breakpoints (or consensus breakpoints) in multiple tumor samples of DNA copy number aberration data. C-KS identifies genomic locations around which breakpoints tend to accumulate more frequently than expected by chance and assigns to each a significance z-score. The algorithm takes as input the locations of breakpoints of all tumors in the cohort and uses a Gaussian kernel for obtaining a moving-average-like statistic reflecting breakpoint abundance along the genome. Then, by means of a permutation scheme, significance z-scores are estimated for each genomic location. Interesting regions are reported, that yield z-score larger than a certain threshold, typically 3.

The authors applied the C-KS algorithm to a number of datasets, that we also show in table 3.1. In the table, we also included the average number of breakpoints identified per tumor as well as the total number of breakpoints identified by the C-KS algorithm.

### Exploring the genetic and epigenetic properties of the consensus breakpoints

We investigated the genomic and epigenomic properties of the consensus breakpoints reported by C-KS. In Figure 3.42 we compare the overlap of consensus breakpoints with CpG islands in all cancer datasets. For each dataset, we compared to two reference sets: a randomly generated control set (using EpiExplorer’s built-in algorithm), shown in blue in the figure, and the set of all breakpoints from all cancer datasets taken together, shown in gray in the figure. The comparison to a random reference shows a clear enrichment in the overlap with CpG islands (with the exception of the glioblastoma dataset). Such enrichment has been reported previously (Abeyasinghe et al., 2003). However, since one can argue that the enrichment is due to the biased selection of array probes that are mostly located within promoters and genes, we repeated the analysis after excluding the regions overlapping with gene promoters and re-evaluated the overlap with CpG islands (result not showed here). The enrichment still holds. This result leads to the hypothesis that the consensus breakpoints, meaning those breakpoints that are likely to play an important role in cancer progression, tend to colocalize with CpG islands, which general are highly functional regions, subject to dynamic regulation in various tissues. Thus, breaking the



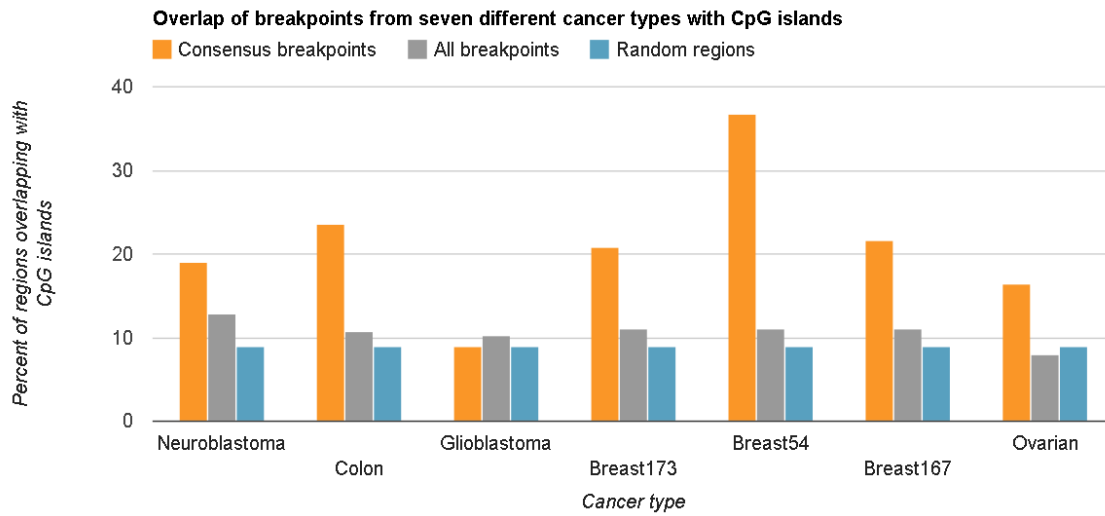


Figure 3.42.: The diagram shows the overlap of consensus breakpoints(orange), all other breakpoints(gray) and randomly selected control set(blue) in seven different cancer cohorts with CpG islands

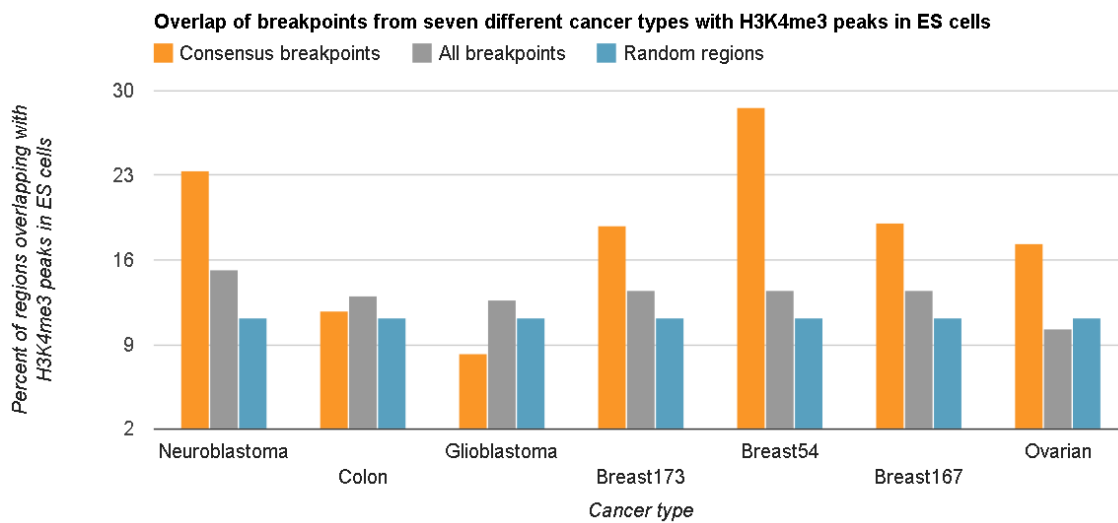


Figure 3.43.: The diagram shows the overlap of consensus breakpoints(orange), all other breakpoints(gray) and randomly selected control set(blue) in seven different cancer cohorts with H3K4me3 peaks in ES cells

DNA within a CpG island with biological function, likely disturbs its function leading to further irregularities in the cell. Next, we investigated the properties of the consensus breakpoints with focus on histone modifications. We started by inspecting H3K4me3, a histone modification that we already showed in section 3.4.1 to be highly associated with CpG islands and especially with unmethylated CpG islands (see Figure 3.24). Not surprisingly, we observe that the consensus breakpoints are enriched in H3K4me3 peaks when compared to all breakpoints and to random regions (see Figure 3.43).

We also inspected another histone modification, commonly associated with repressive

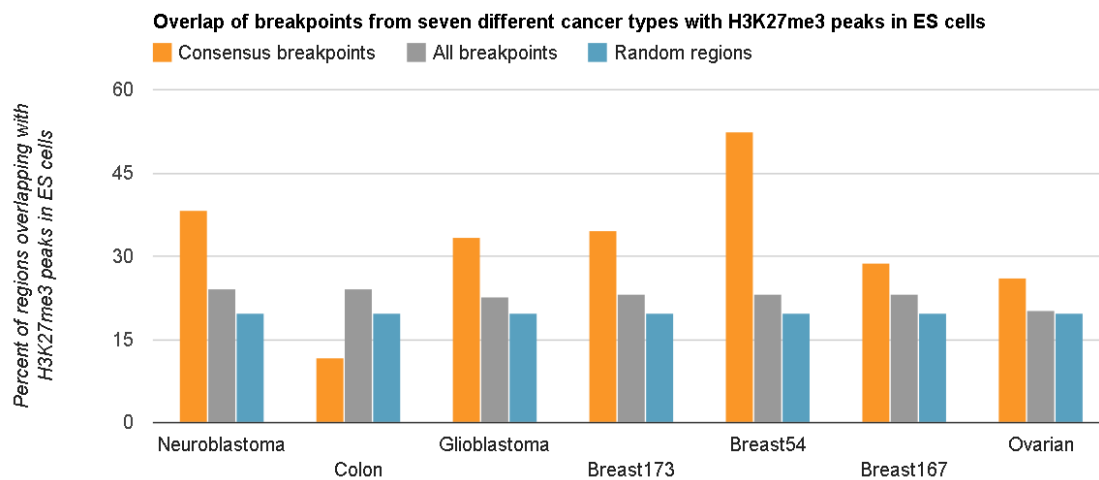


Figure 3.44.: The diagram shows the overlap of consensus breakpoints(orange), all other breakpoints(gray) and randomly selected control set(blue) in seven different cancer cohorts with H3K27me3 peaks in ES cells

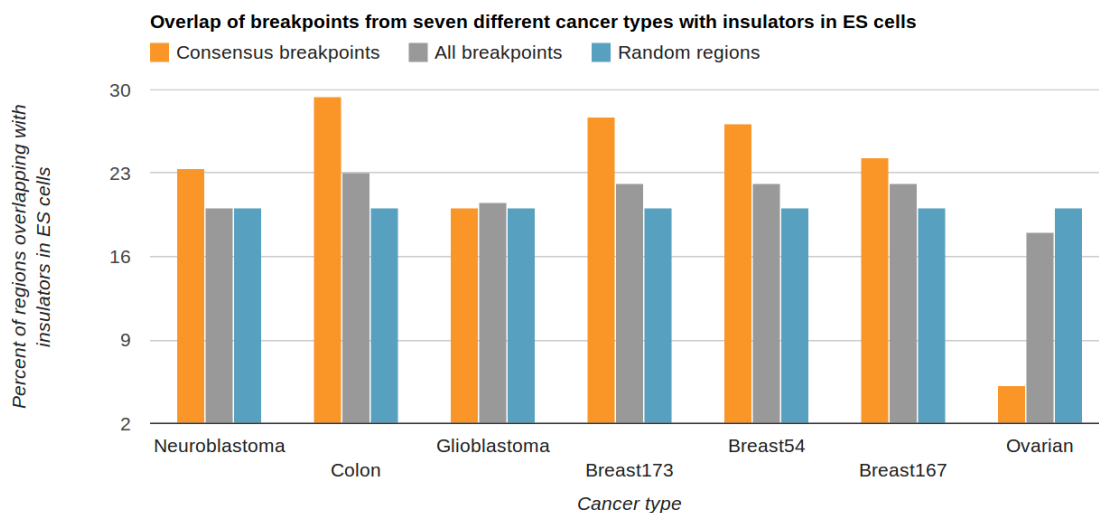


Figure 3.45.: The diagram shows the overlap of consensus breakpoints(orange), all other breakpoints (gray) and randomly selected control set (blue) in seven different cancer cohorts with insulator regions in ES cells

functions and also often observed together with H3K4me3, namely H3K27me3. It marks dynamically regulated CpG islands (more discussion in section 3.4.1). In Figure 3.22, we already demonstrated how the H3K27me3 histone modification tends to target CpG islands in ES cells. Again, We observe an enrichment in the consensus breakpoint regions when compared to individual breakpoints and to randomly selected regions (see Figure 3.44).

We also investigated the overlap with insulators. Figure 3.45 indicates slight enrichment of insulator regions overlapping with consensus breakpoints, in all cancer cohorts except ovarian. The enrichment is an interesting finding, because insulators have been linked to cancer before. For example, the function of the CTCF insulator protein is often disrupted

in cancer, e.g. by hypermethylation of its binding site ((Feinberg and Tycko, 2004)).

Whereas many studies show that the location of the DNA breakpoint in fusion transcripts is critical (eg. BCR-ABL fusion in chronic myelogenous leukemia), little is known about the biological relevance of the breakpoints associated with DNA gain or loss. It is often believed that the oncogenes or tumor suppressors located within the aberrations are responsible for tumor progression, whereas the location of the breakpoint is not essential. This hypothesis is probably true in most cases, as data show that start and end locations of recurrent aberrations may vary greatly. However, many tumors display local aberrations with tightly aligned breakpoints, which suggests that the local structure of the chromatin ‘forces’ the breaks to occur within certain regions by hindering DNA repair (Soria et al., 2012). The EpiExplorer analysis that we performed supports the hypothesis that the DNA locations of tightly aligned consensus breakpoints have interesting biological properties.

In the previous sections 3.4.1, 3.4.2 and 3.4.2, we used EpiExplorer to analyze and explore associations within and between reference genome annotations. However, the analysis above underlines the power of EpiExplorer to be used to draw interesting hypotheses in a disease-oriented study with cohorts of samples.

#### 3.4.4. EpiExplorer performance evaluation

EpiExplorer’s distinguishing feature is the ability to perform a broad range of genome-scale analyses within seconds, thus enabling live exploration, visualization, summarization and interactive filtering of large genomic datasets. Our use of multiple filtering and iterative refinement has important similarities with the concept of faceted search, which is a widely studied paradigm in information retrieval (Hearst, 2009; Tunkelang, 2009). It critically depends on the speed with which complex search queries can be handled. In EpiExplorer, we achieve the necessary runtime performance by using the CompleteSearch engine (Bast and Weber, 2007), which has originally been developed for semi-structured text search in large document repositories. Through creative use of prefix indexing, CompleteSearch provides native support for advanced search features such as query autocompletion and database-style JOIN operations, and has been shown to outperform more standard approaches based on inverted indices (Bast and Weber, 2007). As a result, EpiExplorer was able to complete more than 99% of approximately 4,000 genome-scale analyses performed in the context of the 5hmC case studies in less than two seconds (see Table 3.2).

#### Performance evaluation

Table 3.2 summarizes EpiExplorer’s runtime performance and resource consumption for its five default region sets, as well as for the user-uploaded set of 5hmC hotspots. The preprocessing time needed to annotate and index user-uploaded datasets is usually on the order of minutes to hours (depending on the size of the region set); but it has to be performed only once when a set of genomic regions is uploaded into EpiExplorer. The size of the resulting index structure is typically in the order of few hundred megabytes. Once an index structure has been created, it takes very limited resources for the EpiExplorer server to perform analyses on the corresponding region set. We evaluated the performance of EpiExplorer by measuring the CompleteSearch response times on thousands of queries that were executed during the preparation of the main publication. For every region set,

Dataset	Putative enhancers	CpG islands (specific)	Transcription start sites	Gene promoters (-5kb to 1kb)	5hmC hotspots (Szulwach et al.)	Genome-wide tiling regions (5kb)
Number of genomic regions	1,762	27,638	36,655	36,655	82,221	616,093
Preprocessing time (h)	0.2	0.8	0.9	0.9	1.5	17
Search index size (MB)	11	145	122	127	240	962
Mean query time (s)	0.02	0.06	0.12	0.13	0.2	0.8
95th percentile query time (s)	0.07	0.34	0.5	0.57	0.64	3.2
Percent queries completed in $\leq 2$ sec	100%	99.9%	99.7%	99.1%	99.1%	88%

Table 3.2.: EpiExplorer’s response time and memory footprint across thousands of actual user analyses.

we measured the average query time, the time in which 95% of queries were processed, and the percentage of queries that required less than 2 seconds (Table 3.2). The results show that the average query time for each region set is consistently below 1 second, and that 95% of all analyses even for the largest region set completed in less than 4 seconds, which makes the dynamic exploration of datasets via EpiExplorer a continuous and interactive process for the users.

### 3.4.5. EpiExplorer usage statistics

#### Overview of the first year

Below we present statistics on the usage of EpiExplorer during the first year from publication that indicate its relevance to the bioinformatic community (Table 3.3). EpiExplorer

	Total	Per day
<b>Analysis performed</b>	51,453	141
<b>Dataset sessions</b>	1,603	4.4
<b>Custom datasets computed</b>	1,293	3.5
<b>Paper views</b>	7,594	20.8

Table 3.3.: Overview statistics of EpiExplorer’s first year.

provided more than 50 thousand analyses. These average to around 140 analyses per day. These are distributed into an average of 4.4 dataset sessions per day (a dataset session is defined as at least 5 analyses computed for a specific dataset on a specific day). If we inspect the distribution of analysis by months (Figure 3.46) we notice that most months the number of analyses are between 4,000 and 8,000.

We also inspected the interest in the different genome assemblies that EpiExplorer offers (see Figure 3.47). In the first months after the release of the service we observed similar

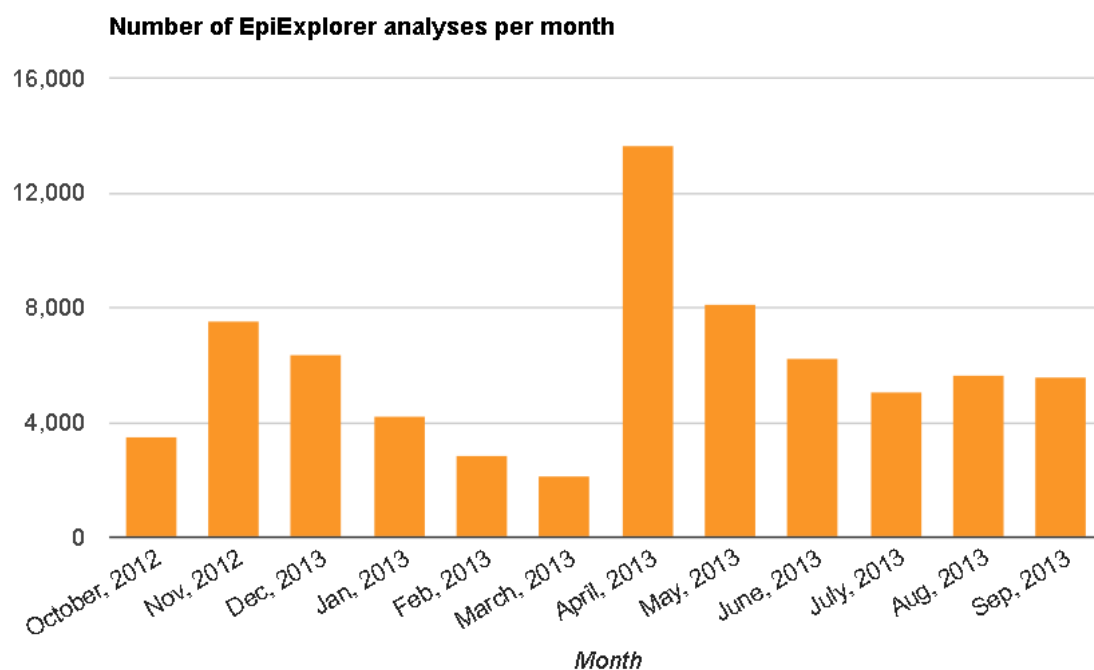


Figure 3.46.: Number of EpiExplorer analysis per month in its first year

interest in hg18 and the latest human genome assembly – hg19. Since the beginning of 2013 we observe a predominant usage of hg19. The analyses on the mouse genome are much more infrequent, with occasional peaks.

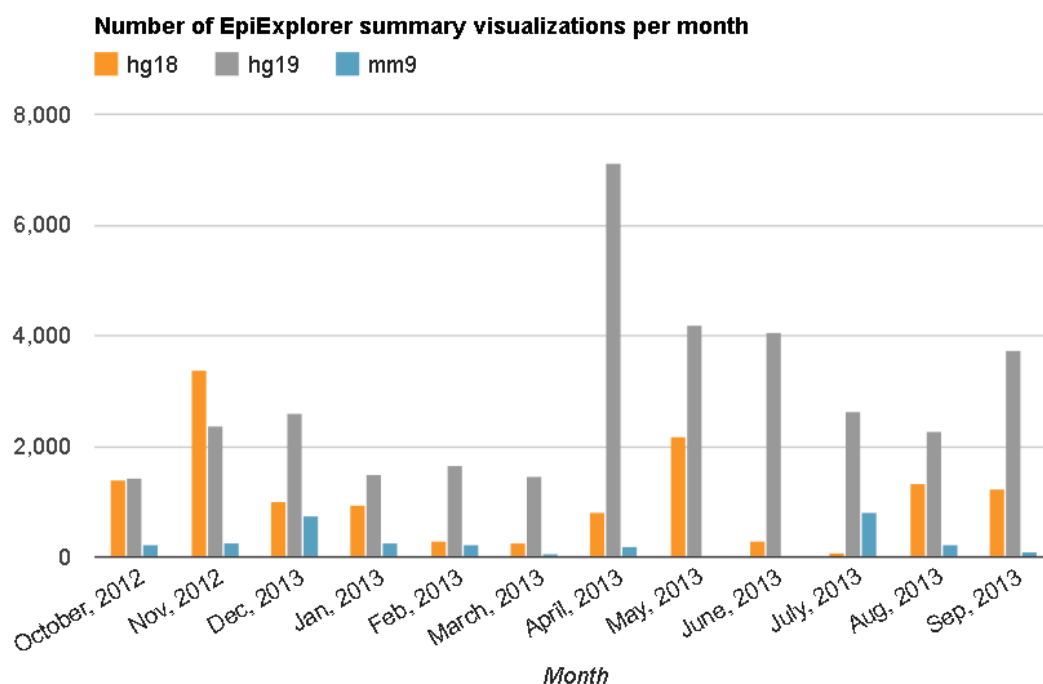


Figure 3.47.: Number of EpiExplorer analysis per genome per month in its first year

As for the different annotations that EpiExplorer offers (see Figure 3.48), we observe the summary views dominating, as is the purpose of these summary views, with the analysis on the histone peaks and chromatin state segmentation following in second and third place. During the first year almost 1,300 custom datasets (uploaded by users) were processed.

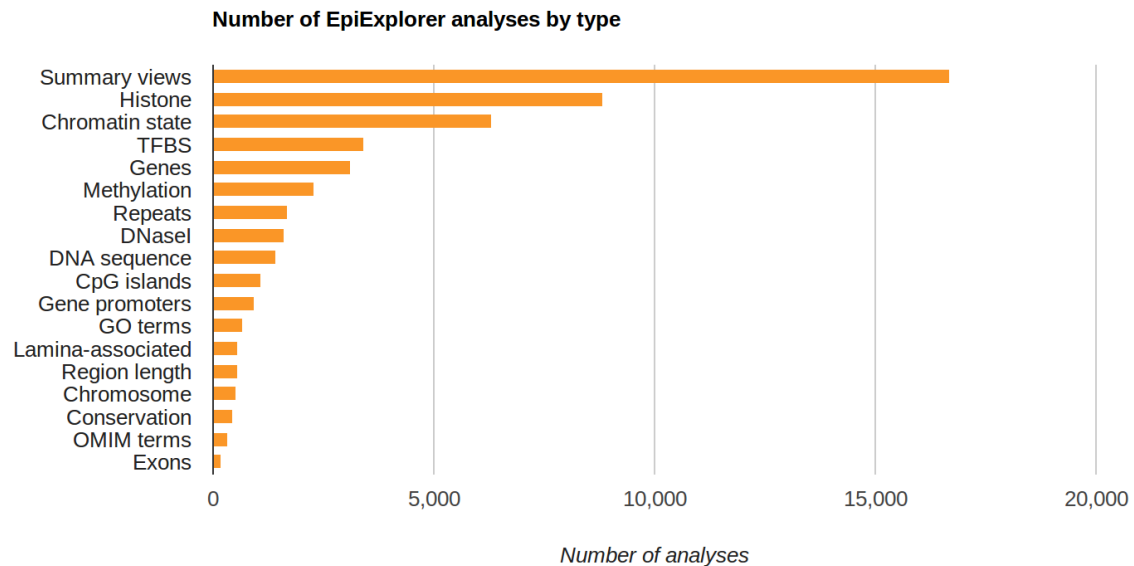


Figure 3.48.: Distribution of performed EpiExplorer analyses for different epigenetic annotations

These average to 3.5 custom datasets per day. Over the year, we observe a consistent interest in computing custom datasets (see Figure 3.49). Averaging 3.5 custom dataset computations and around 140 analyses per day, the EpiExplorer service has proven its value to the biological and bioinformatic community.

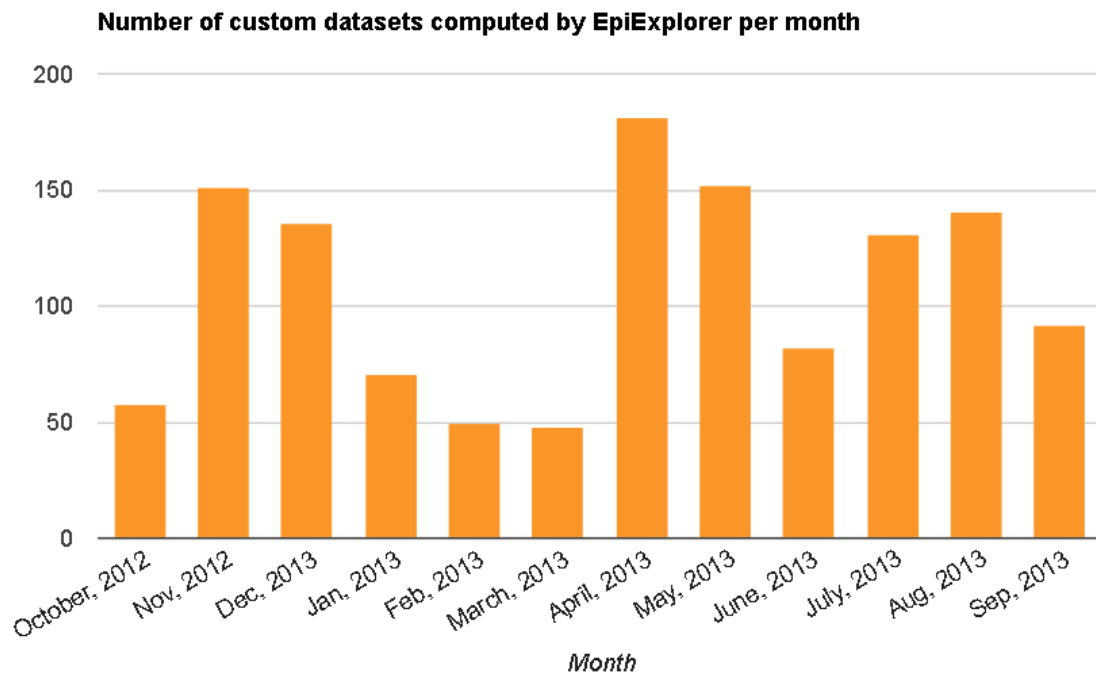


Figure 3.49.: Number of computed custom datasets per month during EpiExplorer's first year

### 3.5. Conclusions and outlook

Toward the goal of interactively exploring large epigenomic datasets, EpiExplorer borrows key concepts from interactive web search. In contrast to genome browsers, which implement browse-and-zoom navigation similar to that of map viewing software, EpiExplorer was inspired by the interactive filter-and-refine workflow of web search engines: Most web searches start broadly (for example, with the search term 'epigenetics') and are then refined iteratively (for example, with the additional terms 'bioinformatic', 'software' and 'tools') until relevant websites show up among the top hits. EpiExplorer supports the same kind of exploratory search by making it easy to dynamically filter genomic region sets and by providing instant feedback in the form of graphical results summaries. Just like web search engines EpiExplorer is highly fault-tolerant, and it enables users to change any aspect of an analysis (for example, thresholds or filtering criteria) at any time without having to repeat previous steps.

The interactive nature of such analyses depends on fast response times, as any delay tends to inhibit the creative act of live data exploration. For this reason, we designed and optimized EpiExplorer to complete complex genome-wide analyses in seconds, rather than the minutes or hours that are the norm for existing genome analysis toolkits (for example, Galaxy (Goecks et al., 2010), Genomic HyperBrowser (Sandve et al., 2010) and EpiGRAPH (Bock et al., 2009)). This level of runtime performance was achieved by utilizing an indexing algorithm that was originally developed for text search. We anticipate that this design principle of EpiExplorer – to encode complex analyses into ultrafast text search queries – will be broadly applicable for interactive analysis of biomedical datasets (for

example, for annotating disease-associated genotypes and in the interpretation of personal genomes).

EpiExplorer successfully delegates complicated and computation heavy aspects to specialized softwares, such as annotation mapping (to BEDtools), text index building (to the CompleteSearch index builder) and query processing (to CompleteSearch). The in-depth statistical analysis of such hypotheses is left to specialized tools such as EpiGRAPH (discussed in the next chapter) and the Genomic HyperBrowser. We validated the EpiExplorer methodology by reproducing already known discoveries about the epigenetic properties of CpG islands. We illustrated EpiExplorer's utility for interactive data exploration by a case study of hydroxymethylation in relation to public reference epigenome datasets, which recreates and extends results from a recently published paper (Szulwach et al., 2011) in ten minutes of analysis time. We also demonstrated that EpiExplorer can be used in disease-specific studies. Detailed tutorials are available from the supplementary website (EpiExplorer: supplementary information, 2012) emphasizing the ease of sharing EpiExplorer analyses and their reproducibility. Finally, we reported usage numbers from EpiExplorer's first year as a publicly available service during which users requested more than 140 EpiExplorer analysis per day. With these examples in mind, we are optimistic that EpiExplorer is a step toward making large-scale epigenome datasets more useful and readily explorable for researchers with little or no bioinformatic experience. Ultimately, EpiExplorer facilitates quick and iterative generation of hypotheses about the interplay of genetic and epigenetic properties.

### **Limitations of the EpiExplorer approach and software**

Using CompleteSearch as a text-search basis for EpiExplorer brings not only benefits, but also limitations. For example, operating with millions of regions sometimes decreases the query response time from milliseconds to seconds. Another drawback is that CompleteSearch does not allow easy sharding (splitting into parts) or updates of its index. A simple task such as exporting the values of an epigenetic property mapped to the set of regions should be easy, as EpiExplorer already computes them during preprocessing. However, these values are indexed and cannot be retrieved easily, therefore EpiExplorer does not support such exports. In order to increment the database of EpiExplorer with annotations of new types, the support team has to manually define the textual mapping (i.e. keywords, values). Frequently, in biological studies, the relation between two epigenetic properties is intuitively pictured by a scatterplot. With EpiExplorer, such visualization is not yet possible, because it is very difficult (within speed requirements) to formulate a CompleteSearch query that returns the necessary values. Via visualizations, EpiExplorer suggests associations between epigenetic properties over a set of regions. However, any sound study should present rigorous evidence that the association holds, typically by statistical hypothesis testing. For now, EpiExplorer does not support statistical validation. Finally, in order to keep the software simple to use and not require advanced querying knowledge, we had to predefine the visualizations and refinements available through the user interface. This limits the types of queries that the system can address (in fact, CompleteSearch can answer a broader set of queries).



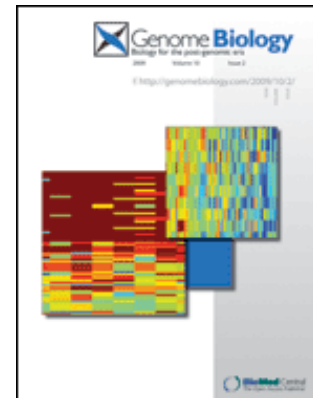
---

**Outlook**

EpiExplorer can be adapted to support the search for genomic regions with specific properties, in the way that text-based engines retrieve documents relevant to certain keywords. Furthermore, it can be extended to searching for regions similar to a region of interest, based on the epigenetic properties within and around the target region. This can later lead to computing a full map of similarity between all genomic regions. EpiExplorer can benefit from integrated hypothesis testing, for example based on open-source toolkits for testing statistical associations. An important future extension of EpiExplorer would be to systematize and standardize its annotation database. Finally, it is possible to provide programming access (API), for example via REST protocol, to the dataset processing and querying service. This will allow EpiExplorer to be easily included in pipelines and will afford a more diverse filtering and querying.



## 4. EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data



Cover of the Genome Biology issue 10-2, 2009 inspired by an EpiGRAPH analysis

In this chapter, we present the methodology, implementation and applications of the EpiGRAPH software toolkit<sup>1</sup>. EpiGRAPH uses statistical methods to automatically identify statistically significant associations between genomic and epigenomic features. These associations are often the result of biological regulation of gene transcription. Understanding the rules of transcription regulation can have a high impact on understanding of disease and advancing progress of medicine.

The rest of this chapter is organized as follows. First, we discuss the background and related work that motivated the development of EpiGRAPH. Then, we present the concept of EpiGRAPH. Next, we describe the methodology supporting the software, discussing in detail the algorithmical, statistical and software engineering solutions. Afterwards, we present three applications of EpiGRAPH to biological studies: **i)** we explore the differences

---

<sup>1</sup>The EpiGRAPH software discussed in this chapter was implemented in close collaboration with Christoph Bock. CB initiated the EpiGRAPH project, conceptualized the software, implemented the front-end, middleware and database components as well as an early back-end prototype. Joachim Büch set up the technical infrastructure. I designed and implemented a substantially enhanced version of the back-end (based on previous work in Halachev (2006)), performed extensive testing and contributed important ideas to all aspects of the project. EpiGRAPH is presented in (Bock et al., 2009; Bock, 2008). In (Bock et al., 2010b), we present how EpiGRAPH can be used together with the Galaxy service. I also performed the studies presented in Sections 4.3.1 and 4.3.2. The study presented in Section 4.3.3 was published in (Feuerbach et al., 2012).

between consistently methylated and consistently unmethylated CpG islands, **ii**) we analyze the association between DNA sequence and DNA methylation in multiple tissues and **iii**) we use EpiGRAPH to analyze the properties of orthologous gene promoters. In the same section, we demonstrate how EpiGRAPH can be integrated into an analysis pipeline with EpiExplorer (see Chapter 3). Finally, we conclude the chapter by discussing the results, challenges and outlook of this work.

## 4.1. Background

EpiGRAPH helps identify and evaluate associations between genetic and epigenetic properties. The discovery of such associations may facilitate the prediction of properties of novel genomic regions for which less experimental data is available. A novelty of EpiGRAPH is that it operates with sets of arbitrary genomic regions. This affords integration of multiple heterogeneous annotations of the genome and epigenome. Bioinformatic software available prior to EpiGRAPH is mostly gene-centric (DAVID (Subramanian et al., 2007), GSEA (Huang et al., 2007)), that can not exploit the rich genome-wide annotations that are stored and available by for example UCSC Genome Browser (Karolchik et al., 2008) and Ensembl (Flicek et al., 2008). Researchers often need to manually process and analyse data. ((Allen et al., 2003; Berry et al., 2006; Karolchik et al., 2008)) Colleagues have noticed the need for automation and proposed a method (Bock et al., 2006) that automatically identifies statistically significant associations between an annotation or a group of annotations with a set of genomic locations. We then extended the method into a fully automated web service, named EpiGRAPH (Bock et al., 2009). EpiGRAPH offers its computational infrastructure and methodology as a public service. Thus, biologists without computational background no longer have to operate directly on data, design statistical studies and create complicated frameworks.

### 4.1.1. EpiGRAPH overview

EpiGRAPH solves the following problem: given a set of genomic regions of interest (*cases*) and a reference set of genomic regions (*controls*), EpiGRAPH maps multiple annotations (*properties*) from a large database of genomic attributes onto the cases and controls. EpiGRAPH then uses statistical testing to identify significantly enriched or depleted properties. It also trains and evaluates machine learning models that can predict whether a given additional genomic region belongs to the cases or to the controls.

EpiGRAPH was developed to address the generic question formulated above, by providing standard statistical and machine learning methodology, which is completely automated and made available as an easy-to-use web tool. Thus, a wide range of analyses are achievable fast by any scientist, without requiring background in bioinformatics. Moreover, using a standardized tool like EpiGRAPH ensures that studies providing a certain type of statistical analysis become easily reproducible and directly comparable.

**Analytical modules.** From the methodological perspective, EpiGRAPH provides four main types of analysis, which we call *analytical modules*. Say that the user is interested in investigating the association between the target property  $T$  and the set of properties  $P_1, \dots, P_n$ . Also, assume that these properties have been summarized as numeric values over

a set of genomic regions  $R_1, \dots, R_k$ , such that each property can be represented as a vector of  $k$  values.

- The statistical analysis module focuses on evaluating univariate associations between some property  $P_i$  and the target variable  $T$ , by means of statistical tests.
- The visualization module provides relevant plots that illustrate the results obtained with the statistical module.
- The machine learning module can be used for building multivariate prediction models for the target variable  $T$  based on subsets of properties  $P_1, \dots, P_n$  and estimating accuracies of such models using cross validation.
- The prediction module provides predictions for the values of property  $T$  on new regions, based on known values of the properties  $P_1, \dots, P_n$  for the region and the model obtained with the machine learning module. The machine learning model is used to train and evaluate models on a training set, while the prediction model is used to predict (and evaluate) the values on the test set.

**Analysis workflow.** A typical analysis with EpiGRAPH (see Figure 4.1) follows these steps:

1. Select data (a set of genomic regions  $R_1, \dots, R_k$  with values of the target property  $T$  for each region).
2. Upload data.
3. Compute attributes (assign values for properties  $P_1, \dots, P_n$  to each of the regions).
4. Analyze statistical dependencies (by means of the statistical module).
5. Generate visualizations (by means of the visualization module).
6. Model  $T$  based on  $P_1, \dots, P_n$  using the machine learning module.
7. Predict  $T$  for a new set of genomic regions.
8. Test predictions externally by wet lab experiments of additional statistical analysis.

**Software.** In the development of EpiGRAPH, we chose established software engineering strategies. For storing and accessing data we use relational databases. For computational efficiency we implemented parallelizable algorithms. In order to easily be able to extend the functionality of the tool with new methods, we opted for a plugin-based software and data architecture. Also, our code is open source.

We generate detailed documentation of each analysis, to ensure reproducibility of the results. The results can be easily shared with colleagues who can inspect and customize the analysis. We also facilitate the export of results in suitable machine-readable format to other tools and custom scripts.

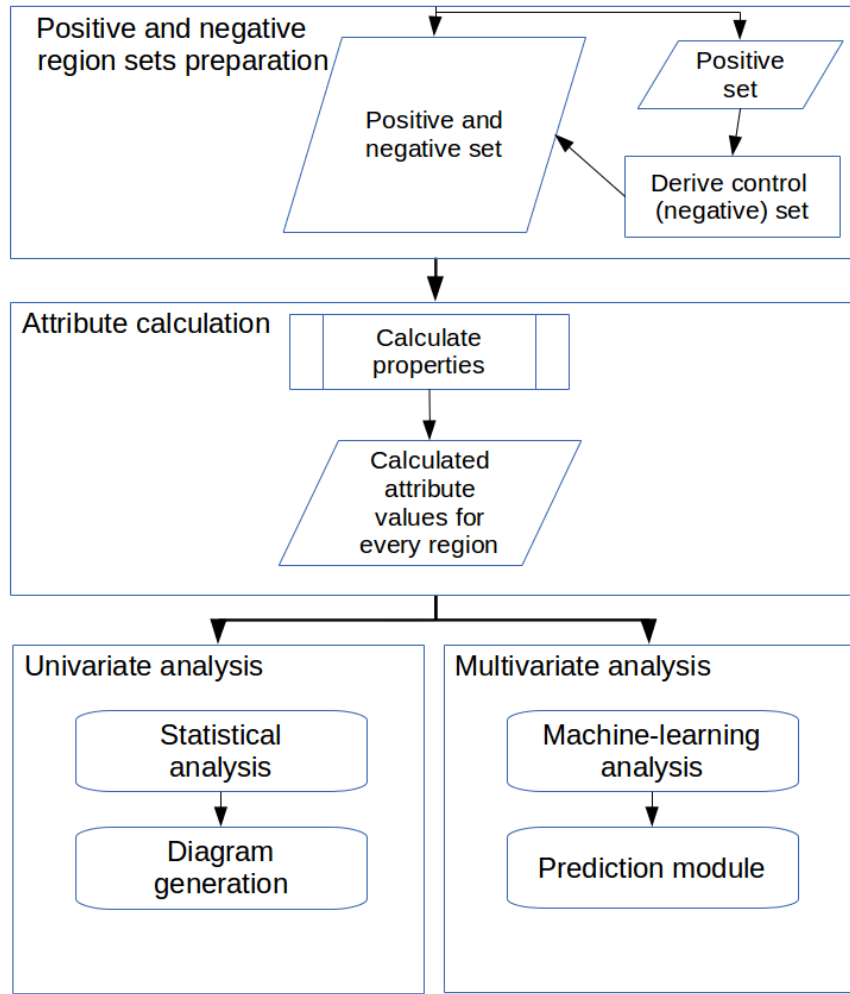


Figure 4.1.: EpiGRAPH analysis workflow. The user prepares two sets of genomic regions (positive and negative). Depending on the set a regions belongs to, it is assigned either 1 (positive) or 0 (negative) value as its target property  $T$ . The user uploads these two sets of regions into EpiGRAPH, where EpiGRAPH computes the representative numeric and categorical values of multiple genomic and epigenomic annotations for each of the regions. Afterwards, EpiGRAPH performs univariate statistical tests and multivariate machine learning modeling to identify and test associations between the properties and the target.

## 4.2. Methods

The efficiency, accuracy and usability of EpiGRAPH relies on a number of methodological decisions. In the rest of this section, we discuss them in detail. We start by describing the algorithm for computing genomic and epigenomic properties for a set of genomic regions. This requires a consistent and automated method for mapping any genome annotation to suitable representative values for any genomic region. Then, we present the statistical and machine learning methods that are at the basis of the EpiGRAPH processes. We con-

tinue with the discussion of the software architecture of EpiGRAPH and specific software implementation details.

#### 4.2.1. Mapping of genomic and epigenomic annotations

##### EpiGRAPH's database of genomic and epigenomic annotations

The EpiGRAPH database is one of the central pillars of the web tool. It supports annotations for four species: *human* (genome assemblies hg18 and hg17), *mouse* (mm9), *chicken* (galGal3) and *chimp* (panTro2). For each of these genomes, EpiGRAPH supports an extensive set of manually curated annotations. These annotations belong to several general groups:

- DNA sequence attributes that describe the base composition of the sequence as well as the distribution of oligonucleotide patterns.
- DNA structure attributes, such as distortions of the DNA helix and predicted solvent accessibility, are inferred from the DNA sequence.
- Repetitive DNA attributes describing repetitive elements, such as transposable elements, tandem repeats and segmental duplications.
- Chromosome organization attributes describe the large-scale functional organization of the chromosomes, such as chromosomal bands.
- Evolutionary history attributes include conservation and local recombination rates.
- Population variation attributes, such as SNPs and microdeletions, describe the variability among individuals, .
- Genes describe the distribution of known and predicted protein-coding genes within the genome as well as the concrete gene parts such as exons, introns, promoters etc.
- Regulatory regions describe putative regulatory regions and functional elements in the genome.
- Transcriptome describe the transcriptional activity, including non-genic transcription.
- Epigenome and chromatin structure describe the chromatin structure and epigenetic modifications, including histone modifications and protein binding.

The numbers of attributes for each genome assembly are listed in Table 4.2.1. For a detailed list of all available attributes see the supplementary resource (Bock, 2009) for EpiGRAPH (Bock et al., 2009).

We collected most of the above mentioned annotations automatically from the UCSC Genome Browser ((Karolchik et al., 2008)). We also added to the database a set of published datasets that we considered to be of high interest to the community. Specifically, we included histone modifications ((Barski et al., 2007)), DNA methylation ((Meissner et al., 2008; Rollins et al., 2006)), regulatory CpG islands ((Bock et al., 2007)), DNA helix structure ((Gardiner et al., 2003)), DNA solvent accessibility ((Greenbaum et al., 2007)),

Attribute groups	hg18	hg17	mm9	panTro2	galGal3
Total	931	911	464	328	414
DNA sequence	189	189	189	189	189
DNA structure	21	21	21	21	21
Repetitive DNA	95	95	91	94	94
Chromosome organization	18	29	15	–	–
Evolutionary history	94	101	–	–	86
Population variation	75	75	–	–	–
Genes	37	60	20	10	10
Regulatory regions	249	259	5	5	5
Transcriptome	49	65	9	9	9
Epigenome and chromatin structure	104	17	114	–	–

Table 4.1.: Numbers of attributes included in EpiGRAPH by type and genome assembly.

tissue-specific gene expression ((Su et al., 2004)), isochores ((Costantini et al., 2006)) and transcription initiation events ((Carninci et al., 2006)).

Furthermore, we offer the user to work with custom annotations, which lends additional value to the EpiGRAPH service. For this purpose, we provide the infrastructure necessary for uploading and processing custom datasets. Custom datasets are available for analysis and can be used just as any other general EpiGRAPH dataset, without any further limitations.

### Computing attributes for genome regions

Most genomic and epigenomic annotations from EpiGRAPH’s database consist of some numeric or categorical property attached to a set of genomic regions. We call these *patch attributes* or *patch annotations*. Let us denote the respective regions by  $A_1, \dots, A_m$ . The user is interested in evaluating the property on her own set of regions  $R_1, \dots, R_k$ . EpiGRAPH extrapolates values for these regions by using some natural and intuitive procedures.

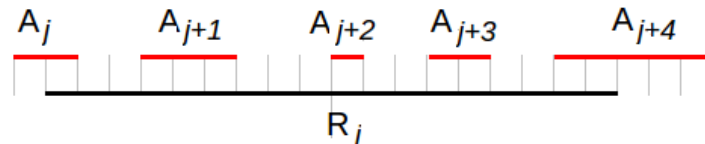


Figure 4.2.: Overlap of a region with a patch annotation. Region  $R_i$  overlaps with regions  $A_j, A_{j+1}, A_{j+2}, A_{j+3}, A_{j+4}$ .

Figure 4.2 shows an example of a region overlapping with five regions belonging to a patch annotation. EpiGRAPH reports three representative values for region  $R_i$ :

- *Overlap frequency*, which is given by the proportion of the region is covered by the annotation, 50% in the example from Figure 4.2.



- *Overlap count*, which is the number of annotation regions overlapping with the targeted genomic region, 5 in the example from Figure 4.2.
- *Overlap average count*, which is the average size the the overlaps,  $(2+3+1+2+5)/5 = 2.6$  in the example from Figure 4.2.

The *overlap count* scores are standardized to a default region size of 1000 in order to be comparable between regions of different sizes. For example, the actual overlap count for the region from Figure 4.2 would be  $(5 * 18)/1000 = 0.9$

In addition, each region from a patch annotations can have an associated numeric score (as in Figure 4.3). EpiGRAPH takes these scores into account and reports additional scores for the target regions, which are meaningful summary statistics over the scores of the overlapping attributes. Specifically, we use a weighted average of the scores of the overlapping annotations, where the weights are the lengths of the individual overlaps. For example, if the scores are as in Figure 4.3, then the score of region  $R_i$  is  $(1 \cdot 1 + 3 \cdot 5 + 1 \cdot 2 + 2 \cdot 1 + 2 \cdot 1)/18 = 22/18$ .

One attribute can have multiple scores, for example the *Bona\_Fide\_CpG\_Islands* attribute proposed in (Bock et al., 2007) reports two scores for every region: the *CombinedEpigeneticScore* and the *OptimizedScore*.

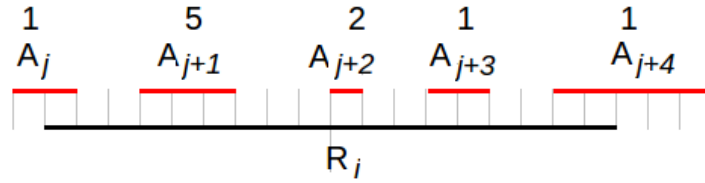


Figure 4.3.: Overlap of a region with a patch annotation with scores. Regions  $A_j, \dots, A_{j+4}$  have additional scores.

Patch attributes can also have categorical values associated with each region. In that case, EpiGRAPH generates the three overlap scores discussed above for every observed categorical value. This is discussed in detail in Appendix B.

Not all attributes in EpiGRAPH's database are patch attributes. Some, such as for example the class of DNA sequence attributes, are genome-wide, meaning that they are not associated to a limited set of regions. We process these attributes separately.

Regarding *DNA sequence*, EpiGRAPH reports for each regions frequencies of patterns of 1, 2 and 4 nucleotides. The counting is not strand-specific. The frequency of a pattern is measured in terms of the number of occurrences of the pattern, divided by the length of the region. However, the frequency does not include information about the distribution of the pattern within the region. For that reason, EpiGRAPH reports helpful statistics such as mean, standard deviation, skewness and kurtosis. In order to compute these values, the region is split into smaller blocks of sequence of equal length and the frequency of the pattern is measured for each. For the resulting distribution of block frequencies, EpiGRAPH computes standard deviation, skewness and kurtosis.

If the user is interested in particular patterns which are not included by default then EpiGRAPH allows her to specify the pattern, whether it is strand-specific or not and which

statistics to be computed (mean, standard deviation, skewness and kurtosis). These will then be computed and used in all standard EpiGRAPH analyses.

*DNA structure predictions* are calculated for a given genomic region by sliding a window of fixed size over the region and comparing the DNA sequence pattern in this window with a set of oligomers with known structure (which is described by numerical score values). For example, the predicted helix structure of all possible octamers has been quantified by a set of six numeric scores: *twist*, *roll*, *tilt*, *rise*, *slide* and *shift* (Gardiner et al., 2003). For each type of score, EpiGRAPH reports one value, which is given by the mean of corresponding scores of all oligomer hits observed while shifting the sliding window over the genomic region. Similar to the pattern frequency attributes described above, we also report standard deviation, skewness and kurtosis.

### Strategies for computing attributes efficiently

Every EpiGRAPH analysis requires automatic calculation of a large number of annotation scores for a potentially high number of regions not only extracted from the rich database of default attributes discussed in the previous part, but also defined and uploaded by the user. EpiGRAPH thus automates one of the most error-prone processes in bioinformatics, extracting and mapping of data. However, this convenience comes at a large computational cost. We use several strategies in order to ensure efficient computation of the attributes. First, we employ multi-threading to parallelize the computations. Second, as we compute scores for consecutive genomic regions, we use caching where possible to reuse already processed resources. Third, even though the default analysis includes all attributes available, we also offer the user to choose a subset of attributes to be computed, should she have a more specific analysis in mind. Last but not least, we ensure that an interrupted calculation (due to external factors) can be restarted and continued from where it failed. Even with these improvements, it is expected that large datasets take a long time to be processed.

From a programming perspective, the computations take place as follows. When a user uploads a set of regions, the EpiGRAPH backend first sorts the regions by chromosome and start position. Next, EpiGRAPH virtually prepares an ordered list of multiple smaller tasks. Each such task involves the computing of the regions scores for a region R and an attribute A. EpiGRAPH starts multiple computational workers (threads) that simultaneously perform such tasks. Each worker receives a task, computes the specified score, then saves the result in a temporary structure and proceeds to the next job. Preserving intermediate results in such structure reduces the amount of data that needs to be stored in memory and allows interrupted computation processes to be easily resumed without loss of data. EpiGRAPH uses additional caching when computing scores for the same attribute and for nearby and overlapping regions as it optimizes the queries it sends to fetch the data needed for the computation of the region result. After all jobs are complete, the results from the temporary structure are assembled and stored in the analysis file. EpiGRAPH provides feedback to the users via its user interface about the stage of completion of their computations as well as an automated email as soon as the computation is complete.

### Derived attributes, neighborhood annotation and control region sets

In addition to standard attributes, EpiGRAPH provides two other types of attributes that ensure a large coverage of biological properties that are frequently explored in studies: *derived attributes* and *neighborhood annotations*.

The user can derive new annotations based on custom formulas involving already existing annotations. For example, a user could add an attribute that computes observed vs expected ratio of the CpG dinucleotides only based on the DNA sequence scores that we provide by default. The user can specify complex analytical functions either automatically via the web user interface or by using scripting language to implement a method that can be plugged in the backend engine.

The users of EpiGRAPH can request annotation scores to be computed not only for the genomic regions given as input, but also for extended regions next to them (eg. 1kb upstream and downstream from the ends of the regions). These regions may contain cis-regulatory functional elements, such as transcription factor binding sites and insulators that are not present in the region itself but influence the biological function of the region. By computing the annotation scores for them and using univariate statistical tests and machine learning modeling EpiGRAPH may help identify cis-acting biological associations.

In many studies, users have only one set of ‘interesting’ or ‘special’ regions, the properties of which they would like to compare to a set of ‘uninteresting’, ‘not special’ regions. EpiGRAPH assists the user in automatically generating a *set of control regions*, in a meaningful way. Specifically, we automatically generate a set which is similar to the target region set in the number of regions and distribution of region lengths. The user can also control the sequence composition of the control set, repeats content and overlap with exons, as we recognize that these are requirements often present in studies.

### 4.2.2. Statistical and machine learning analyses of EpiGRAPH

#### Statistical analysis and diagram generation

As mentioned earlier in this chapter, EpiGRAPH implements statistical methods for identification of attributes which are significantly different between two sets of genomic regions. These two sets of regions are either one set of regions annotated with some binary property, or resulting from one set of regions (target regions), to which a control set has been generated (as discussed in the previous section).

For the purpose of finding discriminating attributes, EpiGRAPH uses well established statistical tests.

*Wilcoxon rank-sum test.* For numeric attributes, the Wilcoxon rank-sum test (also known as the Mann-Whitney U test) is used (Mann and Whitney, 1947). It tests the hypothesis that the distributions of the values of the attributes for the two sets of regions are the same. We choose the Wilcoxon rank-sum test whenever we cannot make any prior assumption on the distributions of the attributes.

Assume given two sets of continuous observations on random variables  $X$  and  $Y$ :  $\{x_1, \dots, x_{n_1}\}$  and  $\{y_1, \dots, y_{n_2}\}$ , respectively. All observations from both sets are sorted in a common ranked list. Assume that  $R_X$  is the sum of the ranks of observations on  $X$  and  $R_Y$  is the sum of ranks of observations on  $Y$ . Under the hypothesis that the distributions of  $X$  and

$Y$  are equal, the following statistic is approximately normally distributed:

$$U = R_X - \frac{n_1(n_1 + 1)}{2}.$$

*Fisher's exact test.* For categorical attributes, EpiGRAPH uses the Fisher's exact test (Fisher, 1922). Assuming that the contingency table of attributes  $X$  and  $Y$  is as in Table 4.2.2, the probability of observing exactly the counts  $a, b, c, d$  is given by the hypergeometrical distribution:

$$P = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{a+b+c+d}{a+c}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!(a+b+c+d)!}$$

	$X_1$	$X_2$
$Y_1$	$a$	$b$
$Y_2$	$c$	$d$

Table 4.2.: Contingency table of attributes  $X$  and  $Y$  for Fisher's exact test.

EpiGRAPH uses a standard default p-value cutoff of 0.05 for all statistical tests, but the cutoff can be altered by the user.

Naturally, EpiGRAPH tests multiple hypothesis in each analysis, leading to the widely known *multiple testing* problem. Specifically, if a large number of independent tests is performed in concert, some null hypotheses can be rejected by chance only. For example, at the 5% level, the chance of rejecting a true null hypothesis is 5%, which means that from 100 tests with true null hypothesis, 5 are expected to be rejected incorrectly. These are false positive instances. To control the expected number of false positives, we perform *multiple testing correction*. For this purpose, EpiGRAPH implements two approaches: the classic and highly conservative *Bonferroni* (Dunn, 1961) correction and the more recently proposed *false discovery rate (FDR)* (Benjamini and Hochberg, 1995). Furthermore, EpiGRAPH reports a table with all significance scores.

To visualize the difference in distributions of a specific numeric attribute, EpiGRAPH automatically generates boxplot diagrams. The user can thus get a visual confirmation of the significant attributes and easily export the result to a scientific paper (in formats such as PDF and PNG).

## Machine learning analysis and prediction analysis

The univariate analysis discussed in the previous section only afford quantification of univariate associations between an attribute and a binary property of a group of regions. A more general and potentially more revealing task is to identify multivariate patterns of attributes that can predict a binary property of a set of regions. For this purpose, we formulate a classification task in which every attribute is a feature and the outcome is defined by the categorical property associated to the set of regions.

EpiGRAPH supports seven classification models: 1) support vector machine with linear kernel; 2) support vector machine with RBF kernel; 3) AdaBoost on tree stumps; 4) logistic regression; 5) random forest; 6) C4.5 tree generator; 7) naive Bayes (Hastie et al., 2009).

These are used from the external library Weka (Hall et al., 2009) with default parameters. To provide a baseline for classification, EpiGRAPH also reports the prediction performance of a trivial classifier that always predicts the majority class.

For estimating the performance of a classifier, EpiGRAPH uses 10-fold cross validation by which the machine learning algorithm is trained on 90% of the regions and tested on the remaining 10%. This is performed for all 10 folds. The cross validation is repeated 10 times with on randomized subsets of the data to capture variation effects. Finally, the performance is reported via several measures, in order to reveal different aspects that may be of interest to the user: percent accuracy, sensitivity and specificity. We also report the Pearson correlation coefficient between the values predicted by the cross validation runs and the real outcome.

EpiGRAPH enables the user to customize the classification task. By default EpiGRAPH uses all attributes from a biological group (defined in Section 4.2.1) as features to train a model, thus training separate models for each group. Then, it reports the association of each group to the response variable. However, the user can choose which attributes to be included for each particular group. She can also merge groups of attributes into a single model. If the set of regions (i.e. samples for the classification) is too large (e.g. in the thousands), a large runtime is expected for training the model. EpiGRAPH facilitates the definition of downsampling scheme, in which the user specifies the maximum number of regions associated with every value of the response. Especially for the cases in which the classes are unbalanced, EpiGRAPH suggests the user to enforce a more convenient ratio between the class cardinalities via downsampling.

Additionally, we implemented (in Java) two extensions to the Weka library. The first ensures that cross-validation sampling is stratified. Stratified sampling enforces the random sampling to preserve the ratio of the possible values of the response variable for each cross-validation sample. The second extension adds information to the standard Weka output and to the EpiGRAPH output, in order to facilitate exporting of results into the ROC package (Sing et al., 2005).

If a classifier has a good performance, the user may be interested in predictions for genomic regions for which the outcome is not known. The prediction module of EpiGRAPH uses 10 models trained on 10 bootstrap datasets drawn from the training set, respectively, to report 10 different predictions for a new region. The predictions are aggregated and the forecast to either class 0 or class 1 is reported. EpiGRAPH also reports the mean and the standard deviation of pertaining to the set of 10 predictions, to give a measure of certainty of the prediction.

### 4.2.3. Software architecture and implementation of the EpiGRAPH service

#### Software architecture

The software architecture of EpiGRAPH is schematically represented in Figure 4.4. The web service consists of three main software components and two logical databases, as well as an XML schema format called X-GRAF that is described in detail later (see Section 4.2.3).

The web-based *frontend* is used to provide an intuitive and interactive user interface that allows to define and customize an EpiGRAPH analysis, submit it to the service, follow its

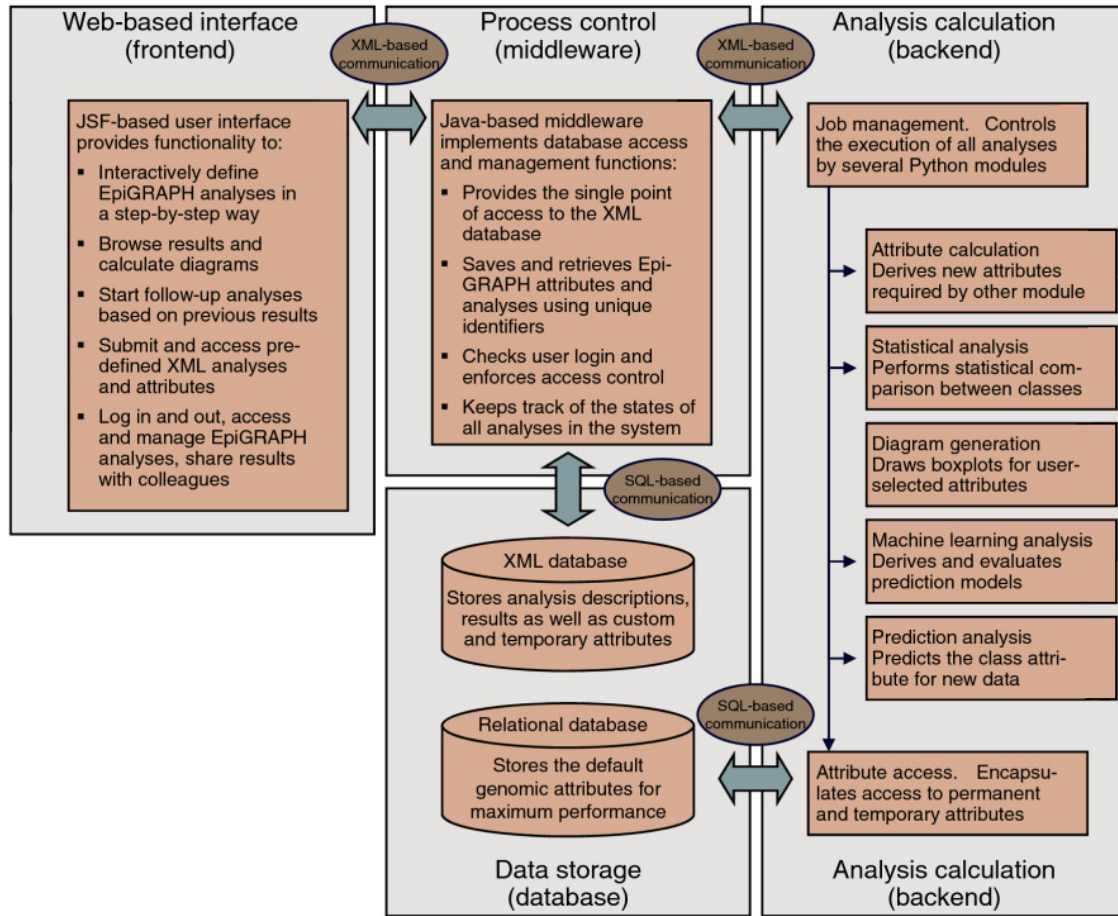


Figure 4.4.: EpiGRAPH software architecture. Figure adapted from Bock et al. (2009)

status and access the results. For this purpose, Christoph Bock implemented the web-based service in Java (Java, 2009) using JavaServer Faces framework, Java servlets and JavaServer Pages.

The user interface is separated from the analysis engine via a *middleware component* that provides access to the analysis and user management. The middleware component provides an interface layer over the access and management of XML analysis. It is implemented as a Java servlet and is accessed from the other components via XML-RPC (Laurent et al., 2001). The middleware provides access to the XML analyses and documentation. These are stored in an Oracle XML DB (Oracle XML DB, 2009), which is an XML extension of the Oracle database.

The main computational component of EpiGRAPH service is the *backend*. It is responsible for all attribute calculations and the follow-up analyses. I implemented the backend in Python (Python Programming Language, 2009), using the R software for statistical calculations and diagram generations (R Project for Statistical Computing, 2009) and the Weka package (Hall et al., 2009) for machine learning and prediction analysis. For attribute calculations EpiGRAPH uses an Oracle database (Oracle Database, 2009) to store the default attributes.

The technologies chosen to implement EpiGRAPH reflect the requirements for the com-

ponent matched to the available solutions at the time of implementation. A simple frontend implementation ensures consistency with a variety of web browsers, which is facilitated by the use of JavaServlet Faces. The middleware needs to support multiple connections and requests to an XML database efficiently, hence Java was used together with specific libraries for Oracle XML database (Oracle XML DB, 2009) and XML processing (van Steensel, 2005; Java Architecture for XML Binding, 2009) were used. The backend implements most of the computational logic and its source code is provided to the public. Therefore, a good choice is the popular Python language, together with the widely-used statistical softwares R and Weka.

### The standalone version of EpiGRAPH

EpiGRAPH also provides two standalone versions to be used and extended by interested researchers. The two versions differ in their usage of a relational database to store default attributes. The database standalone version comprises a backend and a relational database. The no-database standalone version consists only of a backend that expects that all datasets needed for an analysis are provided in the X-GRAF XML format. The no-database standalone version is designed for employing the EpiGRAPH methodology within larger pipelines.

### Internal workflow of an EpiGRAPH analysis

The standard workflow of an analysis with EpiGRAPH runs as follows. First, via the frontend the user sets up the analysis, by selecting data, specifying parameters and choosing methods. The analysis is submitted and sent to the middleware, where after it is verified, it is submitted to the Oracle XML database. The backend regularly checks the XML database via the middleware for analyses waiting to be processed. Each analysis is fetched by the backend. In the most computationally demanding step, the backend maps all attributes required by the analysis and then proceeds with the statistical and machine learning analysis. The processing status of the analysis is regularly updated in the XML database and the user can view it via the user interface. When the analysis is completed, the results or any error notifications are stored in the analysis file in the XML database and the user is notified. She can request the results via the frontend that presents them accompanied by meaningful and intuitive visualizations.

Additionally, the EpiGRAPH backend is accessible via command line. The user can thus submit an analysis in XML format, but can also run more specific computations, such as mapping of a particular attribute. The command-line interface allows to easily submit batches of analyses, as well as to include an EpiGRAPH analysis into a pipelines. Finally, the standalone versions of EpiGRAPH support only the command line interface.

### Extendibility of the EpiGRAPH methodology and data

**Diverse sources of attribute data** Every genomic or epigenomic attribute has to be loaded into the service. EpiGRAPH comprises a plugin-based implementation of the data sources. On one hand, attribute data can be stored in (and loaded from) relational databases, text files or analysis definitions. On the other hand, *calculated attributes* can be defined based on formula-like expressions involving already existing attributes.

The database of default annotations that we support, while sufficient for many analyses, can lack features needed for very specific user needs. The user can expand the set of attributes by defining derived attributes. This can be done either via a genomic calculator interface or by inserting a script-based description directly into X-GRAF XML. For example, the observed vs expected CpG ratio is a score that is often computed and used to determine if a specific genomic region is a CpG island or not. The formula is  $\frac{(\#CpG * N)}{(\#C * \#G)}$  where  $N$  is length of sequence and it can be easily added and computed via the EpiGRAPH genome calculator.

Specifying the parameters of an analysis or attribute calculation directly in an X-GRAF XML file requires advanced scripting skills, but provides more computational options than use of the frontend. Additionally, it is easier to run an analysis with slightly different parameters, by editing the X-GRAF XML file, then by re-introducing it via the user interface. Sharing analyses with other researchers is also easier by sharing X-GRAF files, who can in turn extend, change data and re-run the analysis.

**Adding new statistical, machine learning and visualization modules** By default, EpiGRAPH offers statistical, machine learning, visualization and prediction modules. However, future developments of the service are likely to feature new methods and visualizations. Hence, we made the software implementation easily extensible with new analysis modules.

## The X-GRAF XML format

**Description of the X-GRAF format** EpiGRAPH analyses and attribute definitions are stored in XML files. In order to standardize the format and to ensure the consistent usage of these files between the frontend, backend and middleware, we defined the X-GRAF format. The X-GRAF format implements an XML schema that ensures the correctness of each X-GRAF file and a set of rules that define the meaning of and interaction between the different components (X-GRAF XML Format Documentation, 2009; Illustration of the X-GRAF File Format, 2009). An X-GRAF XML is used for two main purposes: attribute and analysis definitions. The attribute definitions specify how and where the attribute data is stored. The alternatives are: tab-based data directly included in the XML file, link to external databases or links to other attributes and corresponding functional description on how to derive new data. In addition to the data, the X-GRAF format also specifies the semantics of the attribute. Specifically, the biological group to which the attribute belongs is specified, together with some human-readable description. Also, the type of information contained in each column (chromosome start, end, chromosome, scores, categories) must be given, so that the attribute can be interpreted automatically by EpiGRAPH. Similarly, the X-GRAF definition of an analysis has a separate section for each analysis type. Each such section lists the complete settings necessary to specify the analysis, a tracking section that lists the current state of computations of the analysis and finally a results section that embeds the results from the analysis.

**Advantages of using the X-GRAF format** EpiGRAPH stores attribute and analysis definitions, supplementary data and results in the same X-GRAF XML file. There are several



advantages of this approach: reproducibility, documentation, extensibility and ease of collaborative work. By storing the complete specification of an analysis, supplementary data and results in the same file, it contains the full documentation of the analysis and ensures that the analysis can be rerun at any time and will result in the same or similar results (the results may differ slightly if a randomization is employed such as with downsampling for machine learning). The X-GRAF format includes both machine-readable and human-readable sections and thus the XML file together with the X-GRAF rules (X-GRAF XML Format Documentation, 2009; Illustration of the X-GRAF File Format, 2009) can be easily exported as human-readable documentation. The X-GRAF format stimulates collaborative research. Users can share analyses by simply sharing the X-GRAF XML files, which can be easily verified or extended and re-submitted via the web user interface or via the command line interface. Finally, the clear and standardized format allows EpiGRAPH analyses to be easily processed by other software tools.

### 4.3. Applications

In this section we will present three real-world applications of EpiGRAPH. The first study – methylation of CpG islands in promoters – is in fact closely related to the use case that inspired the development of the software. We mentioned this study in the previous chapter (see Section 3.4.1) and we use it as a running example throughout this thesis. Furthermore, the study demonstrates how EpiExplorer and EpiGRAPH can be piped together in order to discover and test interesting associations (see Section 3.4.1). The second use case is concerned with investigating the diversity of DNA methylation across multiple tissues and the varying association with the DNA sequence. In the third use case we utilize EpiGRAPH to perform a cross-species analysis on homologous gene promoters.

#### 4.3.1. Running example: DNA methylation of CpG islands

##### Overview

While the negative genomic correlation between CpG density and DNA methylation has been known for a long time (Bird, 1985), recent bioinformatic studies on predicting DNA methylation have substantially refined our knowledge of which genomic attributes distinguish methylation-prone from methylation-resistant regions. Specifically, it has been reported that the DNA methylation state of CpG islands is highly associated with DNA sequence and repeats (Bock et al., 2006; Das et al., 2006; Fang et al., 2006). These authors used methylation measurements from early experimental technologies in order to construct computational models for DNA methylation of CpG islands.

In section 3.4.1, we used EpiExplorer to extract a set of consistently methylated CpG islands and a set of consistently unmethylated CpG islands. Our results suggested that certain sequence patterns are discriminative regarding the methylation status of the set of regions. However, these observations were mostly visual and not statistically confirmed. Here<sup>2</sup>, we use EpiGRAPH to rigorously evaluate these associations, by means of statistical analysis and machine learning. In the process we demonstrate the analytical power of EpiExplorer and EpiGRAPH working in tandem: EpiExplorer helps to interactively explore various possible associations and identify some for follow up statistical analysis,

EpiGRAPH performs advanced statistical and machine learning analysis to confirm such hypotheses.

We define an EpiGRAPH analysis in which we compare the set of *consistently methylated* CpG islands with the *consistently unmethylated* CpG islands as derived by EpiExplorer. By *consistently methylated* (*unmethylated*) we mean methylated (*unmethylated*) in 7 different embryonic and somatic tissues. This study is closely related to the initial case study that provoked the development of EpiGRAPH in the form of a publicly available service by (Bock et al., 2006). In their work, Bock and colleagues used a prototype version of the EpiGRAPH methodology and database to investigate the differences between methylated and unmethylated CpG islands in 132 islands (data from Yamada et al. (2004)) located on chromosome 21. Additionally, a version of this analysis based on the extended dataset from Yamada et al. (2006) comprising 149 CpG islands is used as a tutorial demonstration of the EpiGRAPH software in (EpiGRAPH tutorial, 2009). The study we present in this section extends these two analyses, by using high-resolution DNA methylation data from multiple tissues to identify the sets of consistently methylated and unmethylated CpG islands (almost twenty thousand CpG islands). Notably, the results reported by the extended analysis are consistent with the limited initial analysis, based on 149 CpG islands from a single chromosome in one tissue.

### **EpiGRAPH analysis confirms significant association between DNA sequence and CpG island methylation**

The EpiGRAPH analysis of the methylation of CpG islands takes the following steps. First, EpiGRAPH computes the values of all default features for each methylated or unmethylated CpG island. This is the most computationally intense part of the analysis as EpiGRAPH has to extract and aggregate the data for hundreds of genomic and epigenomic annotations and match them against almost twenty thousand CpG islands. Then, EpiGRAPH performs a statistical analysis to identify the features that are most discriminatory (in the sense of statistical significance, i.e. smallest p-values) between the positives (consistently unmethylated) and the negatives (consistently methylated) CpG islands. We observe that the most significant feature is *Pat\_CA\_freq* (Figure 4.5). As explained in Appendix B this feature represents the frequency of the strand-unspecific DNA sequence pattern ‘CA’, meaning that EpiGRAPH counted all occurrences of CpA and its reverse complement TpG. Consequently, the results of the EpiGRAPH statistical analysis confirm the suggestion of EpiExplorer (see Section 3.4.1) that TpG and CpA frequency differs between methylated and unmethylated CpG islands. Our finding has biological support. Indeed, it has been shown that erroneous repair of spontaneous deamination causes methylated CpGs to be repaired into CpAs or TpGs.

The next most distinguishing feature is the DNA helix rise as predicted based on the simulation data from Gardiner et al. (2003). As pointed out in the original study ((Bock et al., 2006)), the methylated CpG islands tend to have higher values for the DNA helix rise, indicating unusual DNA structures in methylated CpG islands. However, the DNA helix rise values are calculated based on the DNA sequence, so this association may be mediated by sequence patterns. Among the other significant features we mention the frequency

---

<sup>2</sup>This section reports on an unpublished work performed by the author of this thesis.

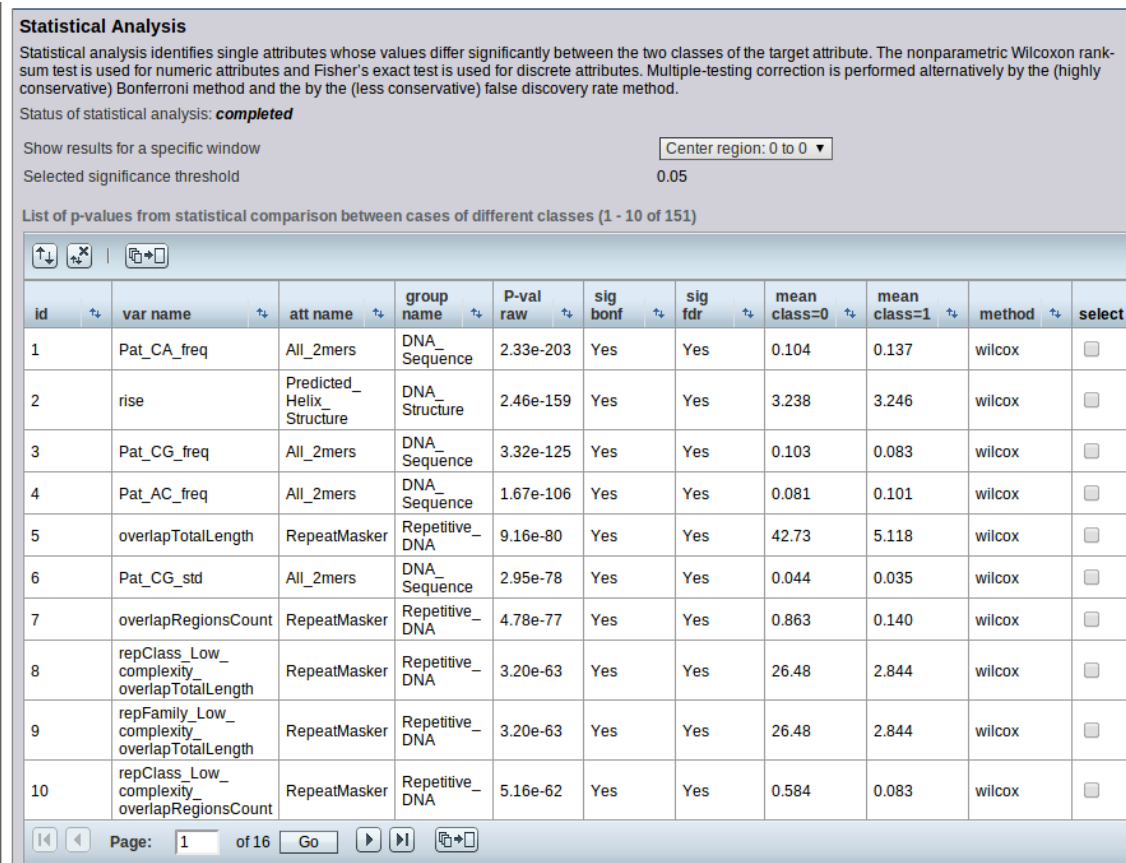


Figure 4.5.: Top ten most significant attributes reported by EpiGRAPH statistical analysis when predicting CGIs with constitutive methylation.

of the CpG pattern as well as its standard deviation. The standard deviation feature is given by the distribution of CpG frequencies obtained by partitioning each region into many consecutive subregions. It has lower values when the distribution of CpGs is similar along the whole region and higher values if certain parts have high CpG frequencies while others are CpG poor. High values of CG\_std are usually indicative of regions overlapping with CGIs. Previously, it has been reported that CpG frequency is a strong predictor of DNA methylation (Straussman et al., 2009; Weber et al., 2007; Bock et al., 2006; Das et al., 2006). Naturally, the characteristics of the distribution of CpGs within a region (eg. the standard deviation *Pat\_CG\_std*) also are associated with the DNA methylation state. 'Low complexity repeats' and 'Simple repeats' are also discriminative features. Enrichment of such repeats in unmethylated CpG islands has been reported previously (De et al., 2013) and has been also suggested by the analysis with EpiExplorer earlier in this thesis (see Section 3.4.1).

Interestingly, the results of our analysis are consistent with a similar EpiGRAPH analysis conducted on the very restricted dataset of Yamada et al. (2006), consisting of only 149 CpG islands on chromosome 21. Figure 4.6 summarizes the results of this analysis.

Figure 4.7 shows the automatically generated box plot that demonstrates the significant difference between the distributions of the CpA pattern in the two subsets of CpG islands.

**Statistical Analysis**

Statistical analysis identifies single attributes whose values differ significantly between the two classes of the target attribute. The nonparametric Wilcoxon rank-sum test is used for numeric attributes and Fisher's exact test is used for discrete attributes. Multiple-testing correction is performed alternatively by the (highly conservative) Bonferroni method and the by the (less conservative) false discovery rate method.

Status of statistical analysis: **completed**

Show results for a specific window Center region: 0 to 0 ▾

Selected significance threshold 0.05

List of p-values from statistical comparison between cases of different classes (1 - 10 of 142)

id	var name	att name	group name	p-val raw	sig bonf	sig fdr	mean class=0	mean class=1	method	select
1	Pat_CA_freq	All_2mers	DNA_Sequence	2.01e-11	Yes	Yes	0.113	0.156	wilcox	<input type="checkbox"/>
2	rise	Predicted_Helix_Structure	DNA_Structure	5.32e-09	Yes	Yes	3.236	3.246	wilcox	<input type="checkbox"/>
3	Pat_GC_std	All_2mers	DNA_Sequence	8.39e-08	Yes	Yes	0.056	0.042	wilcox	<input type="checkbox"/>
4	Pat_AC_freq	All_2mers	DNA_Sequence	3.70e-07	Yes	Yes	0.086	0.116	wilcox	<input type="checkbox"/>
5	Pat_CC_std	All_2mers	DNA_Sequence	3.81e-07	Yes	Yes	0.101	0.077	wilcox	<input type="checkbox"/>
6	Pat_CG_freq	All_2mers	DNA_Sequence	5.23e-07	Yes	Yes	0.082	0.065	wilcox	<input type="checkbox"/>
7	Pat_CG_std	All_2mers	DNA_Sequence	5.38e-07	Yes	Yes	0.055	0.043	wilcox	<input type="checkbox"/>
8	repClass_Low_complexity_overlapTotalLength	RepeatMasker	Repetitive_DNA	9.80e-05	Yes	Yes	41.58	8.266	wilcox	<input type="checkbox"/>
9	repFamily_Low_complexity_overlapTotalLength	RepeatMasker	Repetitive_DNA	9.80e-05	Yes	Yes	41.58	8.266	wilcox	<input type="checkbox"/>
10	rise_skew	Predicted_Helix_Structure	DNA_Structure	2.07e-04	Yes	Yes	0.014	-0.104	wilcox	<input type="checkbox"/>

Page: 1 of 15 Go

Figure 4.6.: Top ten most significant attributes as reported by the original EpiGRAPH analysis on 149 CpG islands located on chromosome 21 Yamada et al. (2006).

### Machine learning analysis successfully models DNA sequence pattern features to predict DNA methylation

Further quantitative information on the strength of the association between DNA sequence and DNA methylation is provided by the EpiGRAPH machine learning analysis (see Figure 4.8). The machine learning module uses cross validation to estimate prediction accuracy when using DNA sequence features to predict the DNA methylation status of CpG islands. The analysis reports a prediction accuracy of 86.6% only based on DNA sequence. We observe strong prediction signal based only on the DNA structure features (83.4%), repeat-based features (71.4%) and gene-based features (67.7%). Interestingly, when we combine all features together the prediction accuracy increases only slightly, to 88.7%, indicating that groups of features different from the DNA sequence add little predictive power. The results on the restricted set of CpG islands on chromosome 21 (see Figure 4.9) confirm the strongest signal coming from DNA sequence (84.5%). Under these experimental conditions, the combination of all feature groups achieves lower prediction accuracy than the DNA sequence itself (81.2%). This can be explained by the small number of samples used with large number of features, which can affect the performance of the model by overfitting.

We also investigate how different combinations of feature groups perform on the same

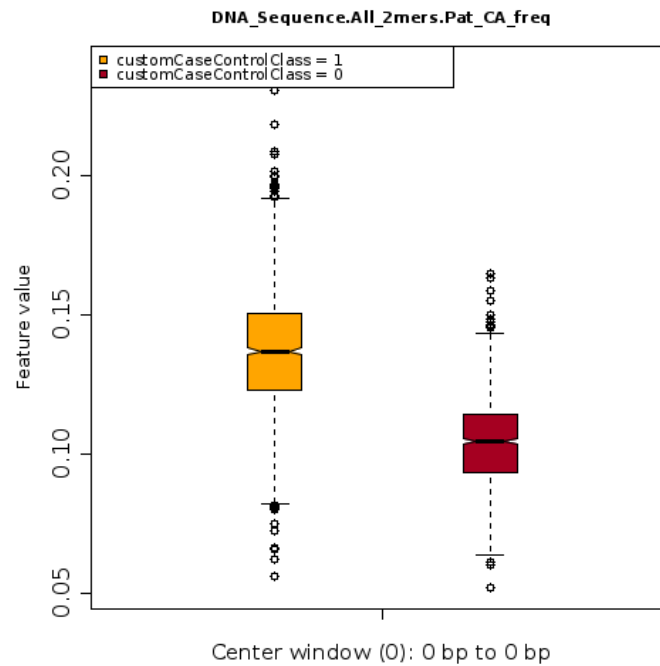


Figure 4.7.: Diagram visualizing the difference of the CpA/TpG distributions

**Machine Learning Analysis**

Machine learning analysis tests how well the two classes of the target attribute can be predicted. The association with different attribute groups is quantified by the prediction accuracy and by the correlation between predictions and true values (averaged over multiple cross-validation test sets). Correlation coefficients above 0.3 can be considered relevant and values above 0.6 indicate strong predictiveness. In addition, a high sensitivity value indicates a low number of false negative predictions.

Status of machine learning analysis: **completed**

Summary of the prediction accuracy for distinguishing between the different classes (1 - 5 of 5)

run	group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	sens	spec	#cases
1	DNA_Sequence	30	svm_linear	0.733	0.004	0.866	0.002	0.866	0.867	2000
2	DNA_Structure	21	svm_linear	0.668	0.005	0.834	0.002	0.818	0.851	2000
3	Genes	10	svm_linear	0.446	0.003	0.714	0.002	0.853	0.575	2000
4	Repetitive_DNA	90	svm_linear	0.413	0.002	0.677	0.001	0.935	0.418	2000
5	DNA_Sequence+DNA_Structure+Genes+Repetitive_DNA	151	svm_linear	0.773	0.003	0.887	0.002	0.889	0.884	2000

[Download Cases List](#)
[Download Performance Summary](#)
[View Settings / Results \(XML\)](#)
[Modify Settings and Recalculate](#)

[View attribute documentation](#) (extensive information on the meaning of the attributes is available from the EpiGRAPH Background page)

Figure 4.8.: Machine learning analysis results when predicting methylation of CpG islands with constitutive methylation. Rows show the performance of various linear SVM models, trained with different sets of features.

prediction task. The results (see Figure 4.10) show that combining the groups ‘DNA structure’, ‘Repetitive DNA’ and ‘Genes’ with the ‘DNA sequence’ group strengthens the predictive power of the model.

All analyses above have been carried out using a linear SVM model. Linearity may limit the performance of the model, as it may not be flexible enough to capture the underlying interplay between features. We repeated the analyses using all machine learning models available in EpiGRAPH. In Figure 4.11 we report the prediction performance of different

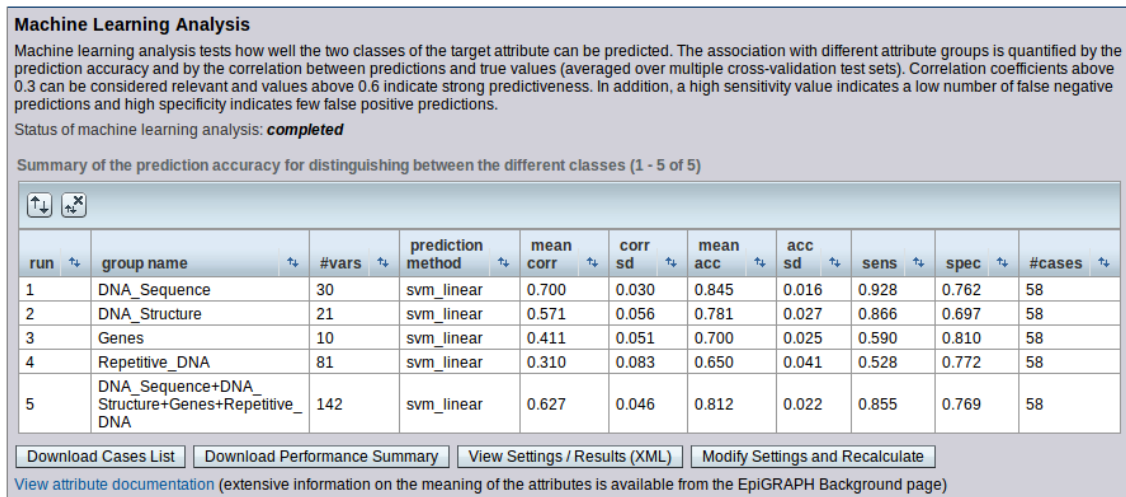


Figure 4.9.: Machine learning analysis results when predicting methylation of CpG islands from Yamada et al. (2006): rows show the performance of various linear SVM models, trained with different sets of features.

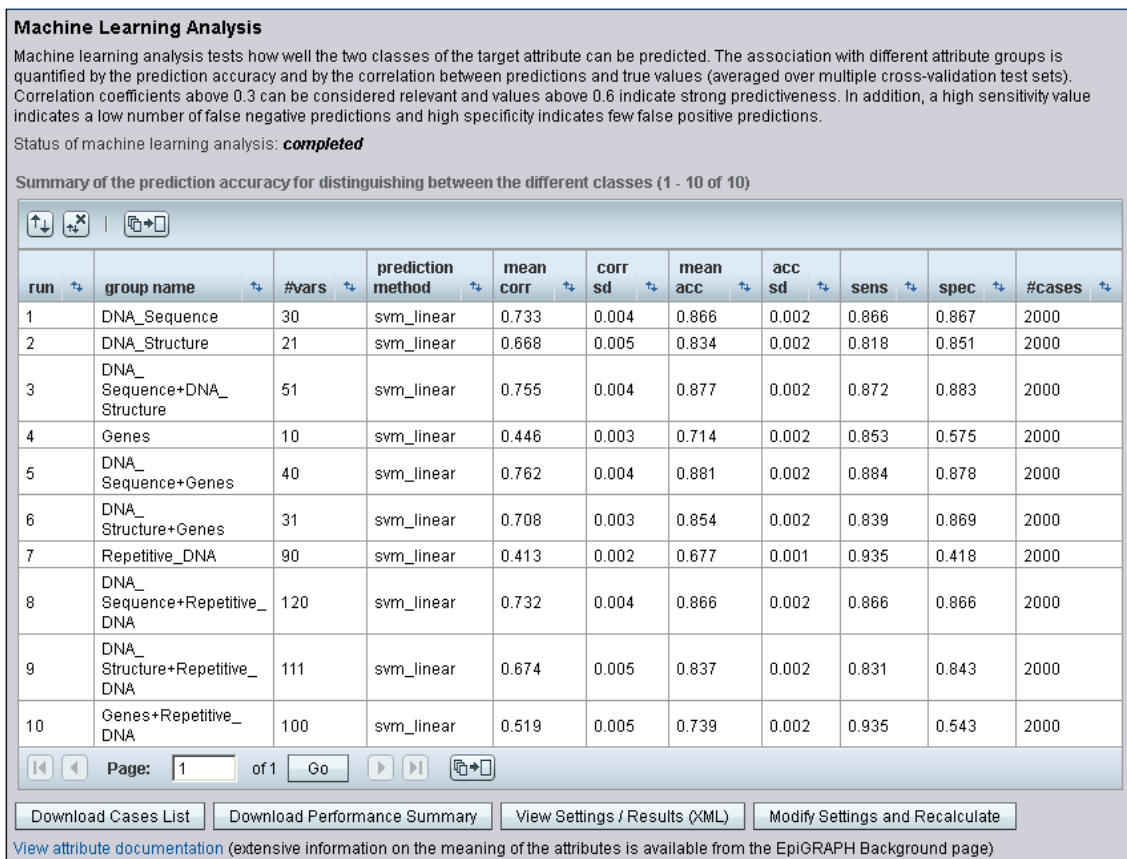


Figure 4.10.: Performance of linear SVM models trained with combinations of groups of features (rows).

models that predict DNA methylation based on DNA sequence. We observe that most of the models report prediction accuracies similar with those of the linear SVMs, between

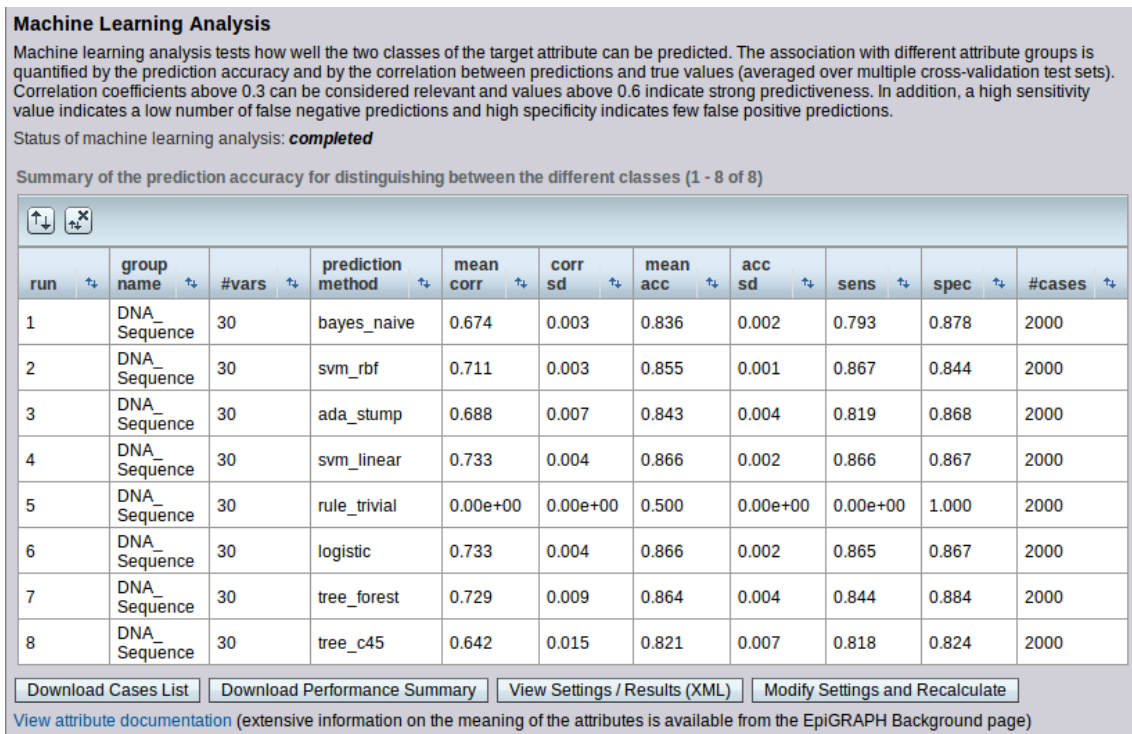


Figure 4.11.: Performance of various machine learning models, trained with DNA sequence features.

84% and 86%. We believe that linear models are already expressive enough to describe the relation between sequence and DNA methylation state, whereas introducing a more complex interplay between features does not improve the prediction models.

## Discussion

In this section, we used EpiGRAPH to test hypotheses suggested by EpiExplorer (in Section 3.4.1). We confirmed the strong association between specific DNA sequence patterns and the DNA methylation of CpG islands. Using EpiGRAPH, we modeled and predicted DNA methylation state of CpG islands only based on their DNA sequence with 86.6% accuracy. However, in this study we did not discuss one important aspect, namely that DNA methylation and histone modifications are tissue-specific, while DNA sequence is not. In the next section, we analyze the tissue-specific aspects of DNA methylation and its association with DNA sequence.

### 4.3.2. DNA methylation in pluripotent cells may constitute an epigenetic ground state

DNA methylation is a key mechanism of epigenetic gene regulation. It is known that DNA methylation dynamically changes during the development of an organism, with two major waves of epigenetic reprogramming occurring in germ cells and during early embryonic development (Reik, 2007). In the previous section, we reported that constitutive DNA methylation patterns in somatic cells can be predicted from DNA sequence with high



accuracy. The result is a rather generic one, therefore we were interested in finding out if the strength of the relation between sequence and methylation status varies among different tissues or stages of cell development. Below we present our findings<sup>3</sup>.

## Motivation

In the previous section, we used EpiGRAPH to evaluate the association between the DNA methylation state of CpG islands and DNA sequence. We formulated a classification task with the target variable being whether a CpG island is consistently methylated or not and reported a prediction accuracy of more than 86%.

With the advancement of next generation sequencing technology, the accuracy and the resolution of methylation measurements improves greatly, allowing for detailed analyses of variation of methylation patterns between individuals (Bock et al., 2008) or across different tissues (Rakyan et al., 2008; Straussman et al., 2009; Meissner et al., 2008). A major limitation of previous published work (Bock et al., 2006; Das et al., 2006; Fang et al., 2006) on prediction of DNA methylation is the focus on single somatic tissues for which sufficient data were available. While it has been argued that DNA methylation patterns are relatively stable between different somatic cell types (Rakyan et al., 2008; Song et al., 2005), there is also considerable evidence that DNA methylation changes in response to developmental clues and environmental influences such as in-vitro culture (Meissner et al., 2008; Mohn et al., 2008). Furthermore, it is well established that DNA methylation follows a complex and dynamic life cycle (Reik et al., 2001; Reik, 2007).

In the view of a dynamically changing epigenome, the question is how to interpret DNA methylation predictions that are derived from a static genome sequence. On one hand, DNA methylation predictions might capture a single stable epigenetic state *encoded* in the DNA sequence. In this case we expect different prediction accuracies between cell types, depending on how similar each cell type's DNA methylation profile is to the sequence-encoded state. On the other hand, the prediction algorithm might be able to identify tissue-specific clues for DNA methylation from training data (e.g. the binding sites of tissue-specific transcription factors), enabling it to predict DNA methylation equally well in multiple cell types.

The available genome-wide DNA methylation data from different tissues enables us to revisit DNA methylation prediction on a large scale. First, genome-scale methods for DNA methylation profiling have been developed and applied across a wide range of cell types (Meissner et al., 2008; Rakyan et al., 2008). We use EpiGRAPH to perform genome-scale DNA methylation prediction for multiple cell types, including pluripotent cells (embryonic stem cells and embryonic germ cells), somatic cells (neural progenitor cells, primary astrocytes, fibroblasts, B cells, T cells, liver tissue, lung tissue, spleen tissue, brain tissue) and in vitro derived cells (neural progenitor cells and astrocytes obtained by in vitro differentiation). We used data published in Meissner et al. (2008); Gu et al. (2010).

In what follows, we train classification models to predict methylation of CpG islands and other genomic regions such as gene promoters, 3' UTRs and others in multiple tissues and we compare the accuracies of these models to better understand the link between static

---

<sup>3</sup>This section reports on an unpublished work. The study was planned together with Christoph Bock. The analysis was conducted by the author of this thesis.



DNA sequence and tissue-specific DNA methylation.

## Methods

Methylation was measured by sequencing genomic DNA from mouse using reduced representation bisulfite sequencing (RRBS). This procedure provides methylation measurements for about 5% of the CpGs, approximately half of which are located within CpG islands. The authors include methylation maps for embryonic stem cells, primary neural cells and other primary tissues. In Table 4.3 we list the available datasets and we enumerate the number of observed CpGs as well as the distribution between methylated (with methylation ratio greater than 0.66) and unmethylated (with methylation ratio less than 0.34). For most of the tissues, the data contains DNA methylation measurements for around nine hundred thousand CpGs that are predominantly unmethylated. Even though, we previously reported that the majority of CpGs are unmethylated, the reported numbers are not surprising as the RRBS technology focuses mainly on CpG dinucleotides within CpG islands (Bock et al., 2010a).

In our analysis, we investigate genomic regions, the biological function of which may be influenced by their methylation state. Our main focus is on (1) gene promoters and (2) bona fide CpG islands (BFCGIs). The latter are defined by (Bock et al., 2007) as DNA regions with  $G+C$  content greater than 50%, CpG observed vs. expected ratio greater than 0.6 and length at least 700bp. We also consider interesting the following sets of regions: (3) middle exons of RefSeq genes with at least 3 exons, (4) 3' untranslated regions (UTRs) of RefSeq genes, (5) putative enhancers (Visel et al., 2009), (6) conserved regions longer than 100bp, and (7) a control set of regions that were selected by randomly sampling from the genome.

For each of the above region sets, we compute DNA methylation levels by averaging over the methylation of each CpG dinucleotide within each region. The scores range from 0 to 1, with low values for unmethylated sites and high values for methylated sites. For the purpose of our study, all regions with scores below 0.34 are considered unmethylated, whereas regions with scores above 0.66 are considered methylated. In order to ensure high confidence in the methylation scores, we also restrict our analysis to regions that contain measurements for at least 5 CpGs. Regions with methylation scores between 0.33 and 0.66 or that contain less than 5 CpG covered by the experimental data are not included in the analyses. Based on these conditions, in Table 4.4 we count the CpG islands for which representative information is available.

We define regions with consistent DNA methylation as regions that have the same methylation state (e.g. methylated) in at least half of the tissues and are not, for example, unmethylated in any of the tissues. All other regions are considered differentially methylated (DMR).

We used EpiGRAPH to predict binary methylation states of genomic regions from attributes derived from DNA sequence. As we discussed previously in this chapter, given a set of genomic regions, EpiGRAPH computes automatically a vast set of DNA sequence pattern features. Based on these features, EpiGRAPH fits a linear SVM classifier that predict the methylation state of genomic regions. Cross-validated (10-fold) prediction accuracies are reported, as well as sensitivity ( $TP/(TP+FN)$ ) and specificity ( $TN/(TN+FP)$ ).

Tissue	Number of CpGs	Proportion of methylated CpGs	Proportion of unmethylated CpGs	Proportion of other CpGs
Embryonic stem cells 300ng (ES_300ng)	736,339	0.204	0.72	0.076
Embryonic stem cells 1000ng (ES_1000ng)	858,720	0.256	0.641	0.103
Lung	796,667	0.278	0.651	0.071
Brain	907,066	0.298	0.618	0.084
Tail-tip fibroblasts (TT)	948,293	0.302	0.611	0.087
Embryonic stem cells 100ng (ES_100ng)	855,232	0.323	0.564	0.113
Spleen	799,697	0.335	0.586	0.079
Embryonic stem cells 30ng (ES_30ng)	916,923	0.343	0.559	0.098
Liver	668,615	0.344	0.59	0.066
Mouse embryonic fibroblast (MEF)	903,946	0.345	0.57	0.085
Embryonic stem cells (ES)	950,703	0.354	0.579	0.067
Embryonic germ line (EG)	952,388	0.354	0.563	0.083
Primary astrocytes passage 2 (Astro_primary_p2)	919,424	0.369	0.558	0.073
Tcell CD4	874,832	0.371	0.577	0.052
Bcell	894,912	0.376	0.576	0.048
Tcell CD8	821,428	0.38	0.568	0.052
ES-derived neural progenitor cells passage 50 (NPC_p50)	1,007,270	0.391	0.541	0.068
Primary astrocytes passage 11 (Astro_primary_p11)	928,225	0.398	0.522	0.08
ES-derived neural progenitor cells passage 9 (NPC_p9)	912,441	0.421	0.482	0.097
ES-derived neural progenitor cells passage 18 (NPC_p18)	921,158	0.425	0.491	0.084

Table 4.3.: Methylation of CpGs in multiple mouse RRBS experiments from Meissner et al. (2008); Gu et al. (2010) sorted by number of methylated CpGs

Balanced sampling was performed prior to model fitting, when the classes were highly unbalanced.

## Results

We trained a support vector machine classification model to differentiate between consistently methylated BFCGs with consistently unmethylated based on their DNA sequence patterns. The estimated prediction accuracy is 92.3% with sensitivity of 94.7% and specificity of 96.6%. Note that this accuracy is higher than the results reported in the previous section for a similar analysis on human DNA methylation. One of the main differences between the analyses is that the current analysis used much more detailed sequence data as all possible 4-mers are also included in the model. Furthermore, the selection of consistently

Tissue	Number of methylated CGIs	Number of unmethylated CGIs	Number of undecided CGIs	Number of CGIs with insufficient data
ES_300ng	174	12,799	337	14,148
ES_1000ng	321	13,013	421	13,703
ES_100ng	356	13,098	483	13,521
ES_30ng	502	13,156	498	13,302
Liver	537	11,275	677	14,969
Lung	627	12,343	388	14,100
EG	692	12,750	601	13,415
ES	709	12,787	593	13,369
Spleen	820	12,189	439	14,010
Brain	910	12,613	391	13,544
TT	1,009	12,636	422	13,391
TcellCD8	1,041	11,839	390	14,188
TcellCD4	1,089	12,201	407	13,761
Astro_primary_p2	1,118	12,026	798	13,516
Bcell	1,121	12,271	369	13,697
NPC_p9	1,639	10,189	2,094	13,536
MEF	1,027	12,242	661	13,528
Astro_primary_p11	1,618	10,996	1382	13,462
NPC_p18	2,384	9,961	1,651	13,462
NPC_p50	2,659	10,478	1,193	13,128

Table 4.4.: Methylation of CpG islands in multiple mouse RRBS experiments from Meissner et al. (2008); Gu et al. (2010) sorted by number of methylated CGIs

methylated cases in the previous analysis was more permissive (consistently methylated in seven tissues), as compared to the current analysis (a region is marked as consistently methylated if it does not have different methylation in any of the tissues). Finally, in the current analysis we use a more restrictive definition of a CpG island, i.e. length of at least 700bp, while in the previous section we used the UCSC annotation that includes CpG islands as little as 200bp in length. The analyses were also performed in difference species. We discuss more about the differences between the mouse and human CpG islands in the Section 4.3.3.

We evaluated the performance of the model trained on the cases with consistent methylation by predicting the DNA methylation state of a test set comprising of all differentially methylated BFCGIs and evaluated the prediction accuracies against the methylation pattern of every individual tissue. The results in Table 4.5 show that the prediction accuracies achieved on the differentially methylated cases are much lower: between 55% (for NPC\_p18) and 75% (for ES). Moreover, we observe higher prediction accuracies in ES and EG cells (75% and 74%) when compared to the rest of the tissues. This suggests that the association between the constitutive DNA methylation and DNA sequence most closely resembles the methylation pattern in embryonic tissues. An alternative interpretation of this result can be that ES cells have the smallest amount of tissue-specific methylation.

Table 4.6 reports the prediction accuracies of models trained to predict the DNA methylation for each tissue separately. The results indicate that the prediction accuracies of most models are above 90%, except neural progenitor cells, for which we report a low 83%. We observe that the methylation in embryonic tissues shows stronger association

Tissue	Prediction accuracy	Number of cases	Sensitivity	Specificity
ES	$75.16 \pm 0.12$	803	76.84	73.98
EG	$74.02 \pm 0.1$	844	77.48	71.83
ES_30ng	$69.9 \pm 0.15$	1127	84.17	65.34
Brain	$69.81 \pm 0.11$	918	70.98	67.97
Bcell	$67.6 \pm 0.1$	960	58.83	69.64
Astro_primary_p2	$66.99 \pm 0.13$	1006	65.26	73.27
Spleen	$66.6 \pm 0.1$	793	59.19	64.15
ES_100ng	$66.24 \pm 0.15$	1037	84.87	62.9
ES_1000ng	$66.2 \pm 0.14$	929	85.43	63.09
MEF	$65.42 \pm 0.14$	911	63.93	70.52
Lung	$65.4 \pm 0.15$	748	66.05	64.74
ES_300ng	$64.25 \pm 0.15$	898	86	62.12
TT	$63.52 \pm 0.12$	978	63.35	63.86
Astro_primary_p11	$62.13 \pm 0.11$	1128	60.05	77.04
Liver	$62.02 \pm 0.11$	628	61.01	63.37
TcellCD8	$59.49 \pm 0.1$	890	57.9	67.72
TcellCD4	$59.27 \pm 0.11$	941	57.55	66.56
NPC_p9	$58.87 \pm 0.12$	973	57.76	71.53
NPC_p50	$58.64 \pm 0.12$	1358	54.92	83.68
NPC_p18	$55.34 \pm 0.12$	1112	54.73	69.99

Table 4.5.: Prediction accuracies on the DMR BFCGI methylation in multiple mouse RRBS experiments based on an SVM model trained on the consistently methylated BFCGIs

with DNA sequence patterns than the methylation of somatic cells. This effect seems to be the result of higher sensitivity or ability to correctly predict methylated cases. This hypothesis has a biological basis: following the early epigenetic reprogramming events, a ground methylation state is restored on the genome and DNA sequence can be one of the main factors that determines the methylation state set immediately after. Later on, as bona fide CpG islands acquire or lose methylation during differentiation, the association between methylation and DNA sequence decreases.

**DNA sequence predictors of DNA methylation state across different tissues** In the previous section, we reported the high prediction accuracy of the models based on DNA sequence patterns when predicting the methylation state of bona fide CpG islands. The statistical analysis of EpiGRAPH quantifies the statistical association between DNA methylation and each of the DNA sequence patterns. In Figure 4.12, we present the properties that were found significantly associated with the DNA methylation in all tissues (the naming conventions for features are explained in Appendix B). First, we observe that the CpA pattern and its reverse complement TpG appear significantly more in methylated BFCGIs. This observation is consistent with our previous findings in Sections 3.4.1 and 4.3.1. The results also indicate that frequencies of *G* and *C*-rich patterns tend to be higher in unmethylated CpG islands. Multiple ‘standard-deviation’ features (‘std’ features), for which higher values indicate irregular distributions of a sequence pattern, have significantly higher values in unmethylated CpG islands. A possible interpretation is that unmethylated CpG islands often have a core with consistent CpG-related sequence patterns, but these patterns

Tissue	Prediction accuracy	Sensitivity	Specificity	AUC
ES_1000ng	$94.5 \pm 0.06$	93.4	95.1	98.7
ES_300ng	$94.4 \pm 0.05$	94	94.6	98.6
ES_30ng	$94 \pm 0.05$	92.5	94.7	98.4
ES	$93.9 \pm 0.04$	91.4	95.1	98.3
ES_100ng	$93.8 \pm 0.03$	92.4	94.5	98.3
EG	$93.6 \pm 0.04$	90.5	95.1	98.2
Brain	$92.7 \pm 0.06$	88.6	94.8	97.3
Lung	$92.6 \pm 0.05$	89.1	94.4	97.3
MEF	$92.2 \pm 0.06$	87	94.8	96.9
Astro_primary_p2	$91.9 \pm 0.06$	86.6	94.6	96.9
Liver	$91.6 \pm 0.06$	86.9	94	96.8
TT	$91.4 \pm 0.06$	85.5	94.3	96.2
TcellCD8	$91.1 \pm 0.09$	84.8	94.3	96.1
Spleen	$91 \pm 0.05$	85.2	93.8	96
Bcell	$90.9 \pm 0.06$	84.6	94	95.8
TcellCD4	$90.6 \pm 0.06$	84	93.9	95.7
NPC_p9	$87.6 \pm 0.1$	76.8	93	93.5
Astro_primary_p11	$87.2 \pm 0.08$	76.6	92.5	92.9
NPC_p18	$84 \pm 0.12$	69	91.4	90.1
NPC_p50	$83.4 \pm 0.06$	68.9	90.7	88.9

Table 4.6.: Prediction accuracies of the BFCGI methylation in multiple mouse RRBS experiments based on SVM models trained for each individual tissue

appear less towards the borders of the CpG island.

We further inspected features that were significantly associated with the methylation profile of at least one, but not with all tissues. The results (see Figure 4.13) show that multiple features appear significant in only few tissues. Interestingly, two major clusters of tissues seem to form: the pluripotent tissues and the somatic tissues. A potentially interesting finding is that the group of features, which are significant only in somatic tissues seem to be highly associated with the TATA-box sequence pattern (TATAAA), possibly indicating that TATA-box BFCGI (usually associated with tissue-specific genes) are unmethylated in somatic tissues, but silenced in embryonic ones.

**Analyzing DNA methylation in different region types** We extended the scope of the study by analyzing DNA methylation in relation to several other groups of genomic functional elements. We evaluated the DNA methylation state of gene body representatives (gene exons and 3' UTR), gene enhancers, conserved regions and a set of arbitrary sampled genomic regions that do not overlap with any of the aforementioned groups. Similarly to the CpG islands analysis, we extracted methylation profiles for these genomic regions and used EpiGRAPH to train classification models that predict DNA methylation states for every combination of genomic region and tissue. The results (see Figure 4.14) confirm that for gene promoters and CpG islands (not the BFCGI definition, but a relaxed definition with length of at least 200 bp), the methylation state is more associated with DNA sequence in pluripotent cells than in adult cells. The same holds for gene body regions (exons and 3' UTR). Interestingly, we get low prediction accuracies for the case of conserved elements and random regions. A possible explanation is that these sets of regions are selected so

Higher in unmethylated CGI		Higher in methylated CGI
Pat_AA_skew	Pat_AAAA_freq	Pat_AC_freq
Pat_AA_std	Pat_AAAT_freq	Pat_CA_freq
Pat_AC_skew	Pat_ACCC_freq	Pat_GA_freq
Pat_AT_skew	Pat_ACTA_freq	Pat_AACG_freq
Pat_AT_std	Pat_AGGG_freq	Pat_AAGA_freq
Pat_CA_skew	Pat_CAAA_freq	Pat_AAGC_freq
Pat_CC_freq	Pat_CCAA_freq	Pat_ACGA_freq
Pat_CC_skew	Pat_CCCA_freq	Pat_ACGG_freq
Pat_CC_std	Pat_CCCC_freq	Pat_ACGT_freq
Pat_CG_skew	Pat_CCCG_freq	Pat_AGAT_freq
Pat_CG_std	Pat_CCGC_freq	Pat_AGCA_freq
Pat_GC_skew	Pat_CCTA_freq	Pat_ATCA_freq
Pat_GC_std	Pat_CCTC_freq	Pat_ATGA_freq
Pat_TA_skew	Pat_CGCC_freq	Pat_ATGG_freq
Pat_TA_std	Pat_CGCG_freq	Pat_CAAG_freq
Pat_plusA_std	Pat_CGGC_freq	Pat_CACG_freq
Pat_plusC_skew	Pat_CTAG_freq	Pat_CAGA_freq
Pat_plusC_std	Pat_CTCC_freq	Pat_CATC_freq
Pat_plusG_skew	Pat_GCCC_freq	Pat_CATG_freq
Pat_plusG_std	Pat_GCGC_freq	Pat_CGTA_freq
Pat_plusT_std	Pat_GCTA_freq	Pat_CGTC_freq
length	Pat TTAA_freq	Pat_CTGA_freq
		Pat_CTTC_freq
		Pat_GAAC_freq
		Pat_GTCA_freq

Figure 4.12.: The table displays all features that were found to be significantly associated with DNA methylation state in all tissues by a p-value threshold of 0.001 after Bonferoni correction. The features on the left have higher values in unmethylated BFCGI, while those on the right have higher values in methylated CGI

that they do not overlap with any gene or CpG island overlap and thus are less often subject to active epigenetic regulation.

## Discussion

In the above analysis, we extended the study that reports DNA sequence being a strong predictor of DNA methylation state. We started by evaluating the performance of the model derived on BFCGIs with constitutive methylation and observed that it predicts best the methylation patterns in pluripotent cells. To estimate the association between DNA sequence and tissue-specific methylation, we modeled the dependency for each tissue individually. We observe strong association in all tissues, but highest performance was obtained for embryonic stem cells and embryonic germ cells. The accuracies are lower in somatic cells and further deteriorate for in-vitro cultured cells. Our results suggest a model in which reprogramming establishes an epigenetic ground state that is largely encoded in the DNA sequence. With the establishment of tissue-specific methylation, the association between DNA methylation and the DNA sequence slowly decreases, until it is reestablished in the following germline transmission. When inspecting the actual sequence patterns, we observe that T/A patterns are indicative of methylated regions and G/C patterns of unmethylated regions. Furthermore, there is a clear group of sequence patterns

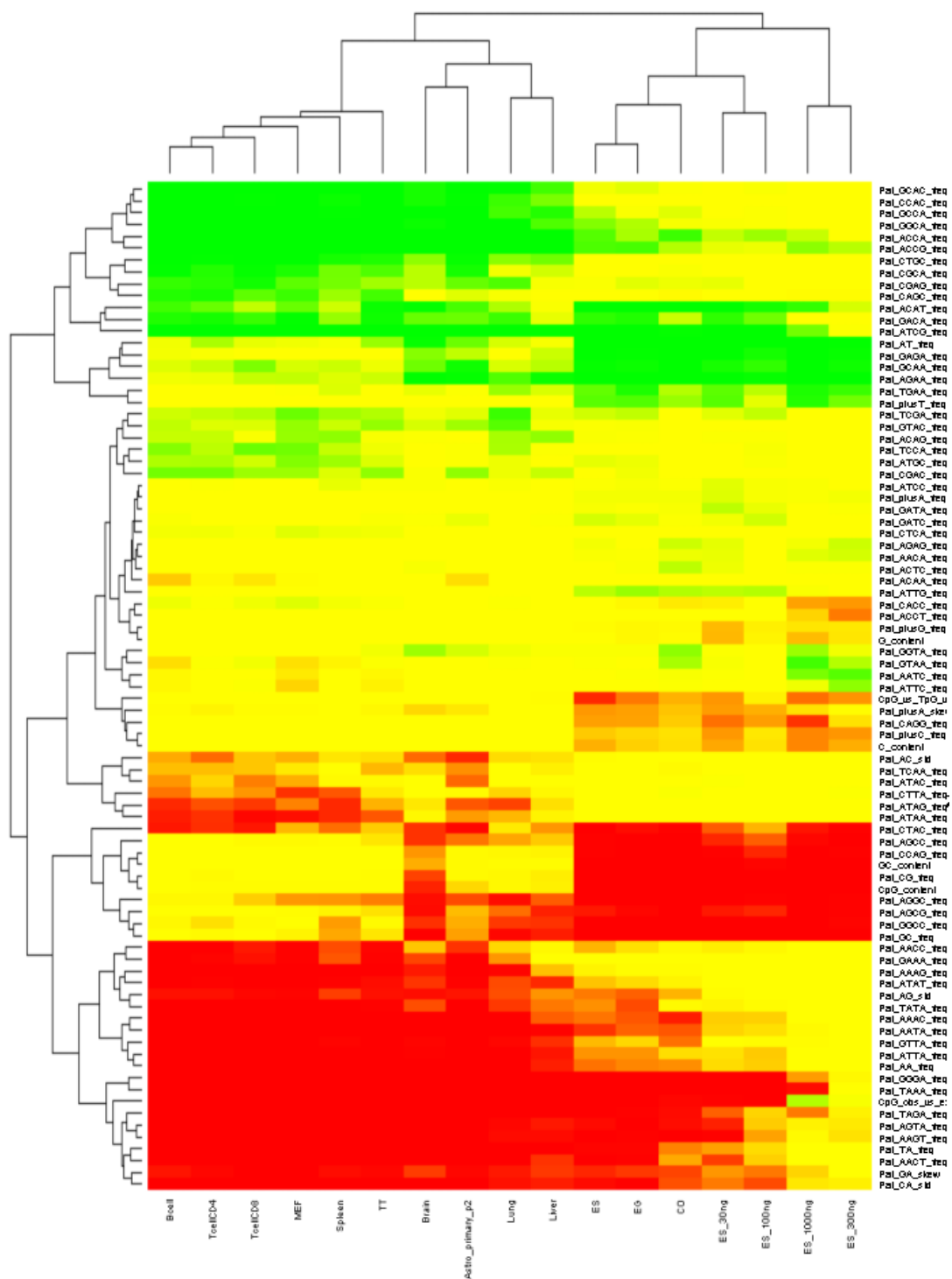


Figure 4.13.: Histogram of DNA sequence patterns that were found significant in at least one tissue. Red color indicates that a feature is significant and the values of the feature are higher in unmethylated regions, green means that the feature is significant and its values are higher in methylated region and yellow indicates that the feature was not found significant in the particular tissue

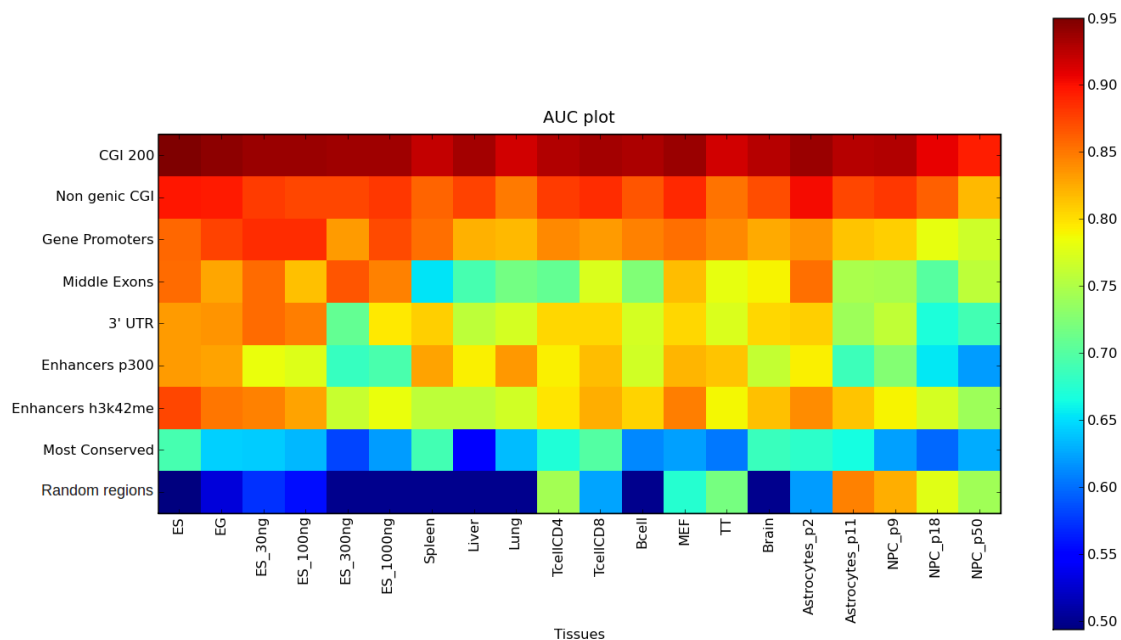


Figure 4.14.: Heatmap representation of the prediction accuracies from machine learning classification scenarios for multiple tissues and region types. Each cell corresponds to a specific setting of genomic region set (rows) and a tissue (column). The color of the cells indicates the prediction accuracy estimation based on a stratified cross-validation of a machine learning classification model based on DNA sequence features.

significant only for the embryonic tissues. Finally, we inspected the DNA methylation of non-genic and non-CGI regions, where we observe little to no predictive associations between the DNA sequence and the DNA methylation.

### 4.3.3. Epigenetics of orthologous gene promoters

This study demonstrates how EpiGRAPH can be used for analyzing epigenomic differences between species. Specifically, we take a look at the CpG islands co-locating with orthologous promoters in human and mouse. The study was part of more extensive work we present in (Feuerbach et al., 2012)<sup>4</sup>.

#### Motivation

In Chapter 2 of this thesis, we described the main epigenetic mechanisms of regulation of gene expression: methylation and histone modifications. Presently, genome-wide and high-resolution assays afford the investigation of epigenetic modifications in various organisms, tissues and in diseased cells. For example, abnormal methylation patterns are associated with a variety of diseases. Such patterns can be used to diagnose functionally deviating

<sup>4</sup>The author of this thesis prepared and carried out the EpiGRAPH analyses reported in this section. These analyses were designed and interpreted together with Lars Feuerbach and his other co-authors in Feuerbach et al. (2012).



cell states (Noushmehr et al., 2010; Figueroa et al., 2010; Yi et al., 2011; Bock et al., 2011). Identifying such associations between DNA methylation states and different diseases and extracting relevant biomarkers is a nontrivial task. It is essential that these associations are studied in appropriate model systems. Studies in mouse are a typical practice. Using other organisms as model is not straightforward, as the conservation of a functional region (e.g. promoter) in a model organism does not imply similar epigenetic regulation machinery. A common studied target is the conservation of gene promoter regulation. Our application aims at analyzing the conservation of epigenetic regulation of promoters between human and mouse. The study takes the following steps.

- We start from a sequence-based classification of human-mouse orthologous gene promoters in two types, co-localizing or not with a CGI. Specifically, we select gene promoters in human and assign them to two different categories, those the orthologs of which co-localize with CGIs in mouse, and those the orthologs of which are not CGIs in mouse. We refer to that as the *CGI-state* of a gene promoter. We use EpiGRAPH to try and identify genomic and epigenomic attributes that differentiate between these categories.
- Next, we involve in the analysis the methylation status of the promoters. We investigate the methylation status depending on overlap with CGIs, in human and mouse. We use EpiGRAPH to find attributes that discriminate between the methylation status of promoters.
- Last, we gather the human and mouse promoters together and use EpiGRAPH to predict the corresponding genome of each case.

### Conservation of CpG islands in orthologous gene promoters in human and mouse

We used the dataset consisting of 3197 manually curated human-mouse orthologous gene pairs reported and analyzed in (Jiang et al., 2007). We used Galaxy ((Goecks et al., 2010)) to determine the gene promoter regions and to match them against the dataset of human and mouse CpG islands. At the end of this preprocessing work (described in detail in (Feuerbach et al., 2012)) we obtained a set of 2910 orthologous gene promoters from mouse and human with their genome coordinates as well as their CGI-state in human and mouse. Then, we used EpiGRAPH to annotate these genomic regions with a large number of genomic and epigenetic features, such as G+C content and histone modifications.

	Mouse CGI promoter	Mouse non-CGI promoter
Human CGI promoter	1820	284
Human non-CGI promoter	152	654

Table 4.7.: Conservation of CpG islands in promoters orthologous in human and mouse

We subsequently partitioned the dataset into different subsets according to promoter type and host species. Specifically, we are interested if promoters overlap with CpG islands (CGIs), in each of the two organisms. Table 4.7 shows that the majority of orthologous

promoters share the same CGI-state. However, there are gene promoters that overlap with a CGI only in one of the species.

By definition, genes orthologous in two species can be traced to a common ancestor. For a pair of orthologous genes in human and mouse, the situation that one of the genes has a promoter overlapping a CGI and the other does not can arise in three different ways: (i) in one species, the CGI has been lost by mutation or genomic rearrangement and alternative regulation mechanisms have become dominant; (ii) in the common ancestor, the gene was alternatively regulated, but then the promoter in one species evolved to be regulated by promoter CpG island DNA methylation ; (iii) the CGI definition fails to correctly classify promoters that are close to violating the relevant constraints. These promoters have been described as intermediate CpG content promoters (ICPs) (Weber et al., 2007). In this last case, even small fluctuations in the general species-specific genome sequence composition can put such a promoter above or below the thresholds of the CGI definition, thus leading to the wrong assumption that a change in biological function has occurred.

A possible mechanism for a loss of CGIs in promoter regions (i) is a slow erosion process that is triggered by increased DNA methylation in the germline followed by subsequent loss of individual CpGs through spontaneous deamination. Such erosion has previously been observed for CGIs in the mouse genome (Matsuo et al., 1993). Here, we investigate if this process is associated with the genomic properties of the promoter and can also be observed at the orthologous human loci (albeit at a slower pace).

### **Features of human CGI promoters predictive of the CGI-state of the orthologous mouse promoter**

The objective of the first analysis is to identify features of human CGI promoters that are predictive of the CGI-state of the orthologous gene promoters in mouse. The two types of promoters distinguished in this study are CGI-associated promoters and non-CGI-associated promoters. Hence, the target variable of the EpiGRAPH analysis is the CGI status of the orthologous promoter in mouse. The features that we investigate for associations include: frequency counts for various DNA sequence patterns, predicted DNA structure, information for overlap with repeats, evolutionary history, population variation, and others.

In the results table of the statistical analysis (Figure 4.15), the features are displayed ranked according to p-value. The statistical test on the frequency of the CpG dinucleotides (Pat\_CG\_freq) reports a very low p-value that remains significant after multiple testing correction, indicating the rejection of the null hypothesis – in our case, that the frequency of CpG dinucleotides in human CGI promoters orthologous to mouse CGI promoters has the same distribution as in the human CGI promoters orthologous to mouse non-CGI promoters. Another feature used to define CGIs – the observed versus expected ratio of CpG within the regions (CpG\_obs\_vs\_exp\_ratio) (defined as  $\frac{(\#CpG * N)}{(\#C * \#G)}$  where  $N$  is length of sequence – is also significantly different between the groups. Also, a more complex measure for CGI strength that integrates the combined epigenetic score for bona fide CGI prediction (Bock et al., 2007) with DNA sequence features shows significant higher values for the human CGI promoters that have matching state in mouse. Furthermore, we notice that the H3K4me3 and H4K20me1 histone modifications are enriched in the

id	var name	att name	group name	P-val raw	sig bonf	sig fdr	Mean (class=c False)	Mean (class=c True)	Stddev (class=c False)	Stddev (class=c True)	method
1	CpG_obs_vs_exp_ratio	GC_and_CpG_Density	DNA_Sequence	1.56e-24	Yes	Yes	0.545	0.644	0.145	0.129	wilcox
2	chromMod_H3K4me3_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	1.23e-23	Yes	Yes	323.8	458.5	186.6	180.2	wilcox
3	Pat_CG_freq	All_2mers	DNA_Sequence	4.38e-22	Yes	Yes	0.038	0.049	0.016	0.017	wilcox
4	CpG_content	GC_and_CpG_Density	DNA_Sequence	4.38e-22	Yes	Yes	0.038	0.049	0.016	0.017	wilcox
5	Pat_CGGC_freq	All_4mers	DNA_Sequence	1.85e-20	Yes	Yes	0.008	0.012	0.005	0.007	wilcox
6	Pat_CG_std	All_2mers	DNA_Sequence	4.73e-20	Yes	Yes	0.042	0.049	0.012	0.011	wilcox
7	chromMod_H3K4me3_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	6.98e-20	Yes	Yes	76.91	131.8	92.21	111.0	wilcox
8	chromMod_H4K20me1_overlapRegionsCount	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	8.91e-20	Yes	Yes	6.980	11.88	10.50	14.54	wilcox
9	chromMod_H4K20me1_overlapTotalLength	NIH_Chromatin_Blood	Epigenome_and_Chromatin_Structure	1.01e-19	Yes	Yes	107.0	167.4	105.8	125.4	wilcox
10	Pat_CCGC_freq	All_4mers	DNA_Sequence	1.27e-19	Yes	Yes	0.010	0.014	0.006	0.007	wilcox

Figure 4.15.: Statistical results of the EpiGRAPH analysis of CGI-state of mouse promoters orthologous to human CGI promoters

human CGI promoters whose orthologs resemble CGI promoters in mouse as well. These post-translational modifications of histones are generally associated with open chromatin and CGIs that are especially enriched for CpGs. However, the experimental data for those histone modifications used in that analysis were obtained only from blood tissues (see (EpiGRAPH attribute documentation, 2009)) and should be interpreted cautiously, as they do not necessarily correlate with histone modification states in other tissues. More precisely, the presence of these marks indicates that a promoter is subject to epigenetic regulation in at least one tissue, but their absence in one tissue does not rule out that the promoter is epigenetically regulated in other tissues. Among the most significant sequence patterns are a measure for the ratio between CpG frequency and the frequency of the spontaneous deamination products TpG and CpG (CpG\_vs\_TpG\_v\_CpA\_ratio) and the CpA/TpG frequency (CA\_freq as search is performed on both strands and thus includes the reverse complement as well). Both values indicate that deamination products are enriched in those promoters that do not have a CGI status in mouse.

As previously mentioned, visual inspection of the data is an important step. The diagram generation module of EpiGRAPH enables the user to inspect the distribution of a feature with respect to the target. The box plot presented in Figure 4.16 indicates that for human CGI promoters the observed versus expected ratio of CpG counts of orthologous to mouse non-CGI promoters is significantly lower as that of orthologous to mouse CGI promoters. Nonetheless, the substantial overlap of the two distributions in the range between 0.55 and 0.65 also indicates that this feature alone does not offer sufficient power to predict whether or not the orthologous mouse promoter of a human CGI promoter also contains a CGI. These observations are quantified by the machine learning analysis, which measures the predictive power of genomic features grouped by biological function (Figure 4.17). We used only the default linear SVM model for this application. Prediction accuracies barely exceed 70%, which suggests that no group of features is strongly predictive if a human CGI promoter of an gene has a CGI promoter in its orthologous gene in mouse.

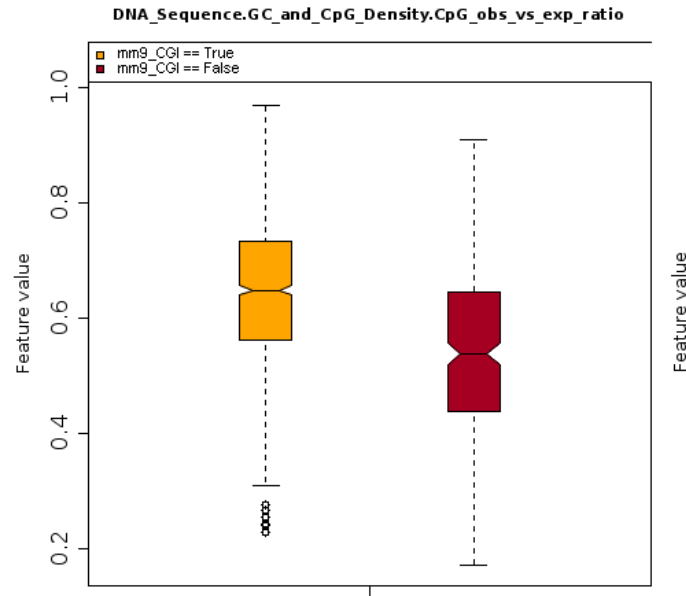


Figure 4.16.: Box plot showing the distributions of observed versus expected ratios of CpG counts of human CpG island promoters that are orthologous to mouse CGI promoters (yellow) or mouse non-CGI promoters (red).

group name	#vars	prediction method	mean corr	corr sd	mean acc	acc sd	sens	spec	#cases
DNA_Sequence	189	svm_linear	0.281	0.020	0.701	0.007	0.852	0.399	849
DNA_Structure	21	svm_linear	0.00e+00	0.00e+00	0.667	1.17e-16	1.000	0.00e+00	849
Repetitive_DNA	197	svm_linear	0.040	0.029	0.657	0.007	0.952	0.067	849
Chromosome_Organisation	24	svm_linear	-0.050	0.016	0.660	0.004	0.989	0.001	849
Evolutionary_History	94	svm_linear	0.083	0.029	0.669	0.005	0.973	0.061	849
Population_Variation	105	svm_linear	6.96e-04	0.019	0.661	0.003	0.984	0.016	849
Genes	43	svm_linear	-0.022	0.008	0.666	3.72e-04	0.998	0.00e+00	849
Regulatory_Regions	279	svm_linear	0.197	0.014	0.677	0.006	0.870	0.293	849
Transcriptome	49	svm_linear	0.011	0.015	0.665	0.001	0.992	0.010	849
Epigenome_and_Chromatin_Structure	107	svm_linear	0.286	0.015	0.712	0.005	0.911	0.313	849
DNA_Sequence+DNA_Structure+Repetitive_DNA+Chromosome_Organisation+Evolutionary_History+Population_Variation+Genes+Regulatory_Regions+Transcriptome+Epigenome_and_Chromatin_Structure	1108	svm_linear	0.242	0.024	0.664	0.012	0.752	0.488	849

Figure 4.17.: Machine learning results of the EpiGRAPH analysis of CGI-state of mouse promoters orthologous to human CGI promoters

In this analysis, we tested and confirmed the hypothesis that human CGI promoters that do not overlap with CGIs at the homologous mouse loci display general properties of ICP-like CGIs (Weber et al., 2007), such as lower frequency of CpG and lower CpG observed versus expected ratio, and furthermore show less evidence for open chromatin, such as H3K4me3 histone modifications. A potential explanation can be that the Takai-Jones CGI definition could be too strict for the mouse genome. The previously mentioned CGI erosion process (Matsuo et al., 1993) has caused loss of CpGs at the boundaries of many CGIs (Jiang et al., 2007) and produced a somewhat shrunken CGI type in mouse. This would primarily affect weaker islands, as those require fewer mutations to be pushed below one of the three thresholds of the definition and as a result not to be considered to be CGIs any more. Hence, not a full change in promoter type explains most of the lost

CGIs, but a slight evolutionary change in their structure that is not reflected in the CGI definition. To test these hypotheses in the context of more epigenetic data, in the next analysis we inspect the DNA methylation properties of the promoters in more detail.

### Analyzing DNA methylation state of orthologous promoters

Here, we analyze the association of DNA methylation and CpG conservation in the context of orthologous gene promoters in human and mouse. For this purpose, we extract methylation information for all orthologous promoters both in human and mouse. We use methylation data obtained from Reduced Representation Bisulfite Sequencing (RRBS) experiments (Gu et al., 2010). RRBS allows for the assignment of a methylation score to every covered cytosine. To obtain a representative methylation score for a promoter, EpiGRAPH averages the methylation scores of the individual CpG sites within this promoter.

CGI promoter	Unmethylated		Methylated	
	Human	Mouse	Human	Mouse
Human and mouse	1746	1759	28	10
Only human	224	94	18	40
Only mouse	14	119	28	3
Neither	34	42	165	137

Table 4.8.: Distribution of methylation data for orthologous promoters visualized by genome and promoter CGI status

We inspect the distribution of the methylated and unmethylated promoters in the different groups of promoters, with respect to their CGI status (Table 4.8). To this end, the methylation information for every promoter is converted from a continuous value between 0 and 1 to a discrete state – methylated or unmethylated. We performed the current analysis, with cutoff values of 0.34 and 0.66 as well as with the stricter 0.25 and 0.75 and we did not observe significant difference in the results (analysis not included).

We observe that promoters in the group ‘mouse CGI promoters orthologous to non-CGI human promoters’ are predominantly methylated in human. In contrast, the corresponding promoters in mouse are predominantly unmethylated. In spite of the relatively small number of cases, it potentially indicates that in human most of these promoters either have lost their ability to be epigenetically regulated or are silenced by DNA methylation in the analyzed tissues. However, in mouse, the majority of these promoters appear to be still epigenetically active, even though they do not meet the CGI criteria.

To set up an EpiGRAPH analysis, we select mouse promoters that are CGI in human but are not CGI in mouse. We exclude all cases that do not have strong methylation scores by adding to the inclusion filter a restriction that methylation score is either less than 0.33 or more than 0.66. The target variable is the methylation status of mouse promoters as a binary value obtained by rounding the methylation score. We analyze the genetic properties of these promoters for significant differences between the methylated and unmethylated promoters. The results (Figure 4.18) indicate that unmethylated non-CGI promoters in mouse have significantly higher frequency of CpG (Figure 4.18A) as well as higher CpG observed versus expected ratios (Figure 4.18B) and lower CpA and TpG frequencies (Figure 4.18C) which indicate CpG decay. The unmethylated non-CGI

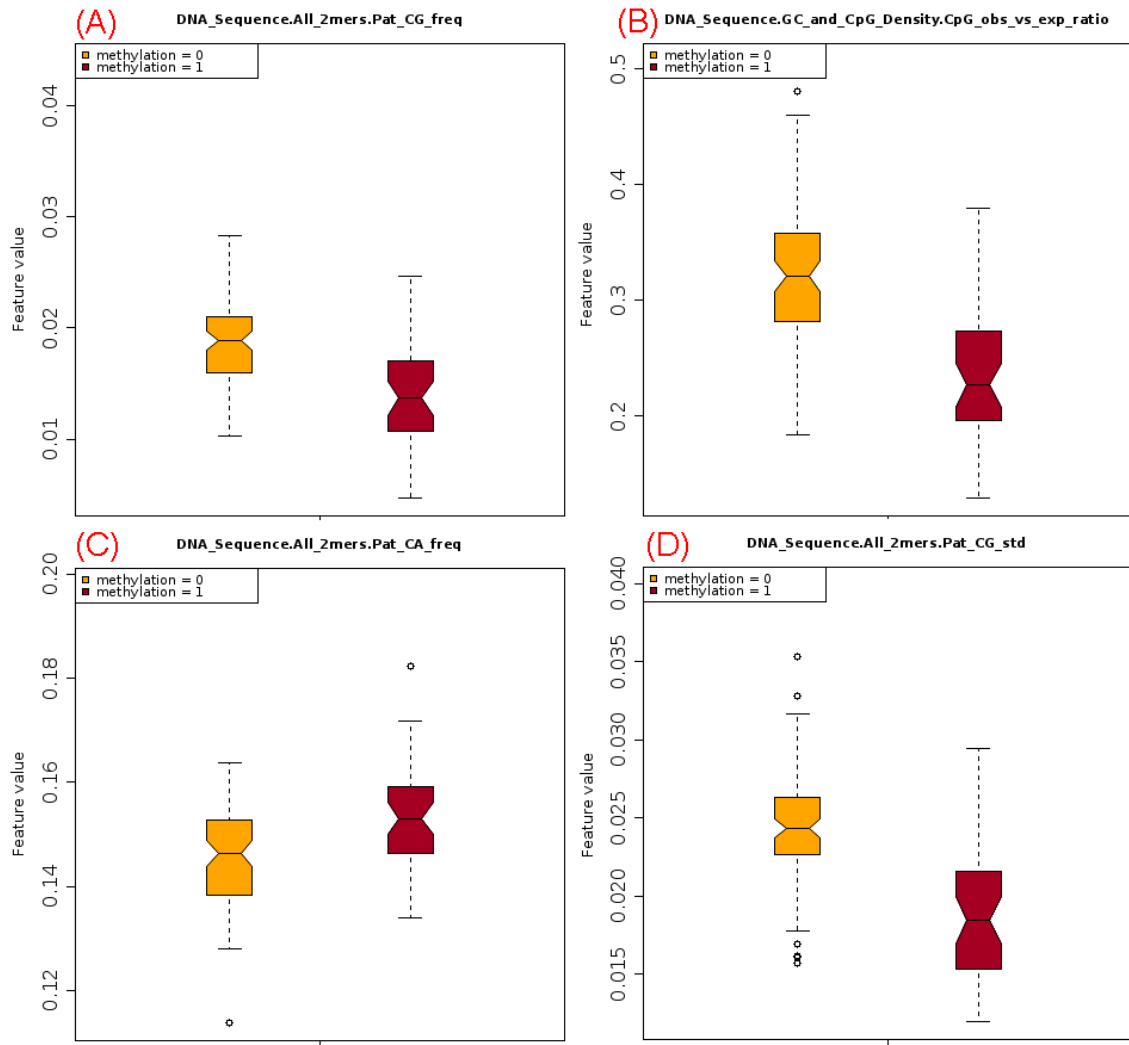


Figure 4.18.: Visualization of some of the most significant features differentiating between methylated and unmethylated mouse non-CGI promoters orthologous to human CGI promoters.

promoters are either protected from this decay or it is considerable slower. Interestingly, the most significantly discriminating feature is the standard deviation of CpG content (CG\_std)(Figure 4.18D). As we discussed previously in this chapter, lower values for a standard deviation of a sequence pattern are associated with a homogeneous distribution of the pattern within the regions (close to a uniform distribution). A possible explanation for the significantly elevated values of this feature in unmethylated non-CGI promoters is the previously described erosion process (Jiang et al., 2007) that starts from the edges of the CGI. Alternatively, the mouse genome may possess smaller CGIs that are somewhat below the minimal length of human CGIs.

These results indicate that among the promoters that lost or never gained CGIs in mouse we observe two different classes. On the one hand, there are the methylated promoters, which apparently homogeneously lose CpGs due to the CpG decay effect. On the other hand, we have the unmethylated promoter type, which represents a shrunken type of CGI

that dropped below the thresholds of the classical CGI definition but still shows many of the classical CGI characteristics. To assess variation in the general evolutionary trends between mouse and human, the next section compares the orthologous promoters with unchanged CGI state. This provides an additional background against which the results from this section can be evaluated.

### Differential analysis of human and mouse promoter traits.

As a follow-up analysis, we test which genomic features are significantly different between human and mouse promoters. For this purpose, we use EpiGRAPH to compute all common attributes for human and mouse promoters (attributes from the EpiGRAPH database available for both human and mouse genomes). We then combine the two sets of promoter regions into one analysis file, to which we add the target variable, called ‘Genome’, which indicates to which organism each promoter belongs to. We then perform a machine learning analysis with EpiGRAPH to identify the differentiating features.

id	var name	att name	group name	P-val raw	sig bonf	sig fdr	Mean (class=c hg18)	Mean (class=c mm9)	Stddev (class=c hg18)	Stddev (class=c mm9)
1	DnaSeq_All4me_Pat_CCG_freq	Common_data	User_Attributes_Attached	0	Yes	Yes	0.010	0.008	0.007	0.005
2	DnaSeq_All4me_Pat_CGCC_freq	Common_data	User_Attributes_Attached	0	Yes	Yes	0.011	0.007	0.008	0.006
3	DnaSeq_All4me_Pat_CGCG_freq	Common_data	User_Attributes_Attached	0	Yes	Yes	0.009	0.006	0.007	0.006
4	DnaSeq_All4me_Pat_GCCC_freq	Common_data	User_Attributes_Attached	0	Yes	Yes	0.014	0.011	0.008	0.005
5	DnaSeq_All4me_Pat_GGCC_freq	Common_data	User_Attributes_Attached	0	Yes	Yes	0.006	0.005	0.004	0.003
6	DnaStr_PreHel_slide_skew	Common_data	User_Attributes_Attached	0	Yes	Yes	-1.879	-2.062	0.452	0.513
7	DnaStr_PreHel_twist_skew	Common_data	User_Attributes_Attached	0	Yes	Yes	-0.396	-0.490	0.144	0.190
8	EvoHis_MulAll_overlapAverageSize	Common_data	User_Attributes_Attached	0	Yes	Yes	4.3e+04	2.7e+04	1.3e+04	1.3e+04
9	EvoHis_MulAll_score	Common_data	User_Attributes_Attached	0	Yes	Yes	0.569	0.437	0.053	0.060
10	RepDna_Rep_overlapAverageSize	Common_data	User_Attributes_Attached	0	Yes	Yes	554.8	384.0	542.1	303.3

Figure 4.19.: List of genomic features that significantly differ between human and mouse promoters.

The results show that the set of features could differentiate almost perfectly (prediction accuracy of 98%) between the mouse cases and the human cases. Furthermore, the features that distinguish most significantly are associated with  $G+C$  content and CpG markers, as well as repeat content (Figure 4.19). We repeat the above analysis only for the True/True group, i.e., promoters that overlap with CGIs both in mouse and human. The results indicate similar predictive power (prediction accuracy of 98%) and show that the human CGI promoters have significantly more CpGs and higher observed versus expected ratio in the context of only slightly higher  $G+C$  content. Furthermore, we notice that the TpG/CpA pattern is more frequent in mouse promoters. Both observations are in accordance with the findings of (Jiang et al., 2007) and indicate that CGIs in mouse have lost CpGs probably due to the CpG decay effect. We also observe significantly higher overlap with repeats for human promoters. As third analysis in this subsection, we compare promoters that are neither CGI in human, nor in mouse. The results point to a number of patterns of A+T-rich patterns indicative for the original genome of the regions. In all cases, the available features could almost perfectly distinguish between human and mouse promoters.

Hence, the EpiGRAPH analyses show that orthologous human and mouse promoters have significantly different genetic and epigenetic features. We showed that the corresponding CGI promoters in mouse have significantly less CpG dinucleotides and enriched products of spontaneous deamination compared to human while orthologous non-CGI promoters differ in mouse and human mainly in their A+T-rich patterns. This indicates that CGI promoters have lost CpG content in mouse compared to the orthologous human promoters.

## 4.4. Conclusions

In this chapter, we present the methodology and implementation of a software toolkit for identifying associations between genetic and epigenetic annotations, called EpiGRAPH. EpiGRAPH supports an extensive database of genomic and epigenomic annotations and offers a powerful backend computation engine that can map these annotations onto sets of regions. The properties of the regions can be analyzed via statistical tests and machine learning models, integrated into the software. The EpiGRAPH analysis is provided as a public web service and is accessible via a user interface. Thus, the users do not need to have programming abilities in order to process their data. In this chapter, we also demonstrate how EpiExplorer and EpiGRAPH can be used together in a pipeline. We show the performance of the software on a large scale analysis and finally, we go beyond genome borders in a cross-species study of orthologous promoters.

*Limitations and future directions.* The EpiGRAPH annotation database is loaded into an Oracle database. While efficient, the Oracle database requires nontrivial maintenance, complex updates and requires advanced programming skills to implement efficient annotation mapping. Furthermore, the need for database support may have prevented researchers from hosting local EpiGRAPH instances, despite us offering the EpiGRAPH source code and providing standalone versions. In the meantime, while developing EpiExplorer, we delegated annotation mapping to open-source software libraries, such as BEDtools (Quinlan and Hall, 2010) that work directly with annotation BED files. This reduced the maintenance need and increased the speed of dataset computations. Probably, EpiGRAPH can benefit from the same approach, namely to delegate most of the annotations processing and mapping to specialized tools.



## 5. Conclusions

Biologists and the medical community increasingly recognize the power of computational technologies and statistical methods to assist with understanding biological processes, due to the increasing size and quality of the experimental data available. On the other hand, increasing number of scientists with computer background are attracted by the applications in the field of bioinformatics. In the field of genetics and epigenetics, improved sequencing technologies produce high quality experimental data at reasonable cost, providing the prerequisites for development of advanced data visualization and mining tools. In this thesis, we presented two software toolkits and the related methodology: EpiExplorer and EpiGRAPH. They are available as public web services, easy to use by biologists and bioinformatic researchers without requiring programming skills.

With EpiExplorer, we provide a tool for interactive mining and visualization of large epigenomic datasets. EpiExplorer moves away from the classic concept of genome browsers, which focus on a single genomic locus at a time, but instead, it provides overviews on sets of regions. The visualizations provide information on the proportion of the regions that co-localize with specific genomic annotations (e.g. conserved regions), on the distributions of DNA sequence patterns within the regions, the distribution of histone modifications in the vicinity of the regions and many others. These visualizations are requested via the user interface through easy mouse clicks and are provided instantly. The user is not limited to a fixed set of regions, but can interact with virtually all subsets of the dataset she starts with. The qualities of EpiExplorer stem from an innovative combination of the best practices from four data analysis fields. Specifically, EpiExplorer uses the power of text search engines (*web search*) on genomic and epigenomic data (*bioinformatics*) to provide faceting overviews (*e-commerce*) on dynamic intersections of the data (*business intelligence*). During its first year, users uploaded an average of 3.5 custom datasets into EpiExplorer and it answered about 140 analyses daily.

We demonstrated the use of EpiExplorer in a validation study where we reproduced known properties of CpG islands. Then, we inspected the properties of a novel epigenetic mark – 5hmC – where we observed strong association with the H3K4me1 histone modification and co-localization with enhancer elements. Furthermore, we used the ability of EpiExplorer to refine a dataset to select a candidate set that can be further analyzed in an experimental setting. Finally, we demonstrated how EpiExplorer can be used with cohorts of patient data as we looked into datasets representing DNA breakpoints from seven cancer types (identified from arrayCGH data). We observed that recurring DNA breakpoints in these cancers tend to be in genomic locations with functional epigenetic markings, as opposed to the non-recurring breakpoints. These analyses demonstrate how EpiExplorer can be used to generate biological hypothesis that claim non-random associations among two or more epigenomic or genomic properties.

EpiGRAPH, the second toolkit and methodology that we introduced in this thesis, fo-

cuses on providing a sound statistical and machine learning framework for automatically testing hypotheses of associations among genomic and epigenomic annotations on given sets of regions. In this manuscript, I discussed in detail the design and implementation of the general and scalable backend that sustains the heavy computational tasks – training large classification models, assessing feature relevance, generating meaningful reports to the user. We demonstrated how to connect EpiExplorer and EpiGRAPH in a single analysis workflow: EpiExplorer suggests that there is an interesting association between methylation of CpG islands and DNA sequence patterns and EpiGRAPH proves that the association holds, with statistical testing. More generally, we showed that DNA sequence patterns are strongly predictive of the methylation state of CpG islands. We then looked into how the strength of the association changes in different tissue types. Finally, we moved beyond the single organism bounds and analyzed the epigenomic properties of gene promoters orthologous in mouse and human and their association with DNA sequence and DNA methylation.

We provided both EpiGRAPH and EpiExplorer as open-source, publicly available web services. When implementing the tools, we emphasized on reproducible and documented analyses that are easy to share with colleagues. One of the key aspects of these tools is their ability to sustain large databases of epigenetic and genetic maps, the maintenance of which is entirely hidden from the users. In that way, we facilitate complex analyses that otherwise require advanced computational and programming skills. With the expected increase in available data, maintaining easy interaction, instant querying and continuous updates of such integrated databases remains a major challenge to the next generation of epigenetic software.

## 5.1. Outlook

A major challenge in our work was to assemble, maintain and update the databases of EpiExplorer and EpiGRAPH. Nowadays, there are multiple distributed relational or NoSQL databases (mongoDB, Couchbase, SciDB) and search indexes (ElasticSearch, Solr, Lucene) available. These provide solutions that are transparently scalable on multiple servers as the size of the data changes, support easy updates, reliable data storage and are easily deployable in the cloud. Scalable cloud-based storage with integrated indexing and querying services can enable instant access to the enormous datasets expected in the near future in the bioinformatic field. Providing such general solutions for access to biological datasets will save bioinformaticians the large overhead of maintaining local copies of these databases and empower them to focus on answering the biological questions they are interested in.

Finally, we emphasize again the importance of reproducible and verifiable results at also discussed in Goecks et al. (2010). At this moment, most analytical results are shared only as a static view (usually a figure or a table) in a scientific paper. However, we (and a large part of the community) feel that this is not satisfactory. Current technology allows already for better sharing of results. With our tools, we make sure that fellow researchers can seamlessly reproduce and extend our research with their own ideas within minutes. With most of biological datasets being public and most tools being offered as public web services, the bioinformatic community has fewer excuses not to embrace fully open research.

# Appendices



## A. EpiExplorer annotations listing

An up-to-date listing of the EpiExplorer annotations is available here: <https://docs.google.com/spreadsheets/pub?key=0AmGLN6XZ0HmydGNoSi1pVDRmQkg30ERURkh5N09NX1E&output=html&widget=true>.

### A.1. Human genome

#### A.1.1. hg19

Table A.1.: Full listing of the EpiExplorer's annotation datasets for human genome assembly hg19.

Annotation name	Annotation group	Annotation source
Active promoters (GM12878)	Chromatin state segmentation	ENCODE
Active promoters (H1-hESC)	Chromatin state segmentation	ENCODE
Active promoters (HepG2)	Chromatin state segmentation	ENCODE
Active promoters (HMEC)	Chromatin state segmentation	ENCODE
Active promoters (HSMM)	Chromatin state segmentation	ENCODE
Active promoters (HUVEC)	Chromatin state segmentation	ENCODE
Active promoters (K562)	Chromatin state segmentation	ENCODE
Active promoters (NHEK)	Chromatin state segmentation	ENCODE
Active promoters (NHLF)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (GM12878)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (H1-hESC)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HepG2)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HMEC)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HSMM)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HUVEC)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (K562)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (NHEK)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (NHLF)	Chromatin state segmentation	ENCODE
Insulators (GM12878)	Chromatin state segmentation	ENCODE
Insulators (H1-hESC)	Chromatin state segmentation	ENCODE
Insulators (HepG2)	Chromatin state segmentation	ENCODE
Insulators (HMEC)	Chromatin state segmentation	ENCODE
Insulators (HSMM)	Chromatin state segmentation	ENCODE
Insulators (HUVEC)	Chromatin state segmentation	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
Insulators (K562)	Chromatin state segmentation	ENCODE
Insulators (NHEK)	Chromatin state segmentation	ENCODE
Insulators (NHLF)	Chromatin state segmentation	ENCODE
Poised promoters (GM12878)	Chromatin state segmentation	ENCODE
Poised promoters (H1-hESC)	Chromatin state segmentation	ENCODE
Poised promoters (HepG2)	Chromatin state segmentation	ENCODE
Poised promoters (HMEC)	Chromatin state segmentation	ENCODE
Poised promoters (HSMM)	Chromatin state segmentation	ENCODE
Poised promoters (HUVEC)	Chromatin state segmentation	ENCODE
Poised promoters (K562)	Chromatin state segmentation	ENCODE
Poised promoters (NHEK)	Chromatin state segmentation	ENCODE
Poised promoters (NHLF)	Chromatin state segmentation	ENCODE
Polycomb repressed (GM12878)	Chromatin state segmentation	ENCODE
Polycomb repressed (H1-hESC)	Chromatin state segmentation	ENCODE
Polycomb repressed (HepG2)	Chromatin state segmentation	ENCODE
Polycomb repressed (HMEC)	Chromatin state segmentation	ENCODE
Polycomb repressed (HSMM)	Chromatin state segmentation	ENCODE
Polycomb repressed (HUVEC)	Chromatin state segmentation	ENCODE
Polycomb repressed (K562)	Chromatin state segmentation	ENCODE
Polycomb repressed (NHEK)	Chromatin state segmentation	ENCODE
Polycomb repressed (NHLF)	Chromatin state segmentation	ENCODE
Repetitive CNV (GM12878)	Chromatin state segmentation	ENCODE
Repetitive CNV (H1-hESC)	Chromatin state segmentation	ENCODE
Repetitive CNV (HepG2)	Chromatin state segmentation	ENCODE
Repetitive CNV (HMEC)	Chromatin state segmentation	ENCODE
Repetitive CNV (HSMM)	Chromatin state segmentation	ENCODE
Repetitive CNV (HUVEC)	Chromatin state segmentation	ENCODE
Repetitive CNV (K562)	Chromatin state segmentation	ENCODE
Repetitive CNV (NHEK)	Chromatin state segmentation	ENCODE
Repetitive CNV (NHLF)	Chromatin state segmentation	ENCODE
Strong enhancers (GM12878)	Chromatin state segmentation	ENCODE
Strong enhancers (H1-hESC)	Chromatin state segmentation	ENCODE
Strong enhancers (HepG2)	Chromatin state segmentation	ENCODE
Strong enhancers (HMEC)	Chromatin state segmentation	ENCODE
Strong enhancers (HSMM)	Chromatin state segmentation	ENCODE
Strong enhancers (HUVEC)	Chromatin state segmentation	ENCODE
Strong enhancers (K562)	Chromatin state segmentation	ENCODE
Strong enhancers (NHEK)	Chromatin state segmentation	ENCODE
Strong enhancers (NHLF)	Chromatin state segmentation	ENCODE
Transcriptional elongation (GM12878)	Chromatin state segmentation	ENCODE
Transcriptional elongation (H1-hESC)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HepG2)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HMEC)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HSMM)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HUVEC)	Chromatin state segmentation	ENCODE
Transcriptional elongation (K562)	Chromatin state segmentation	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
Transcriptional elongation (NHEK)	Chromatin state segmentation	ENCODE
Transcriptional elongation (NHLF)	Chromatin state segmentation	ENCODE
Transcriptional transition (GM12878)	Chromatin state segmentation	ENCODE
Transcriptional transition (H1-hESC)	Chromatin state segmentation	ENCODE
Transcriptional transition (HepG2)	Chromatin state segmentation	ENCODE
Transcriptional transition (HMEC)	Chromatin state segmentation	ENCODE
Transcriptional transition (HSMM)	Chromatin state segmentation	ENCODE
Transcriptional transition (HUVEC)	Chromatin state segmentation	ENCODE
Transcriptional transition (K562)	Chromatin state segmentation	ENCODE
Transcriptional transition (NHEK)	Chromatin state segmentation	ENCODE
Transcriptional transition (NHLF)	Chromatin state segmentation	ENCODE
Weak enhancers (GM12878)	Chromatin state segmentation	ENCODE
Weak enhancers (H1-hESC)	Chromatin state segmentation	ENCODE
Weak enhancers (HepG2)	Chromatin state segmentation	ENCODE
Weak enhancers (HMEC)	Chromatin state segmentation	ENCODE
Weak enhancers (HSMM)	Chromatin state segmentation	ENCODE
Weak enhancers (HUVEC)	Chromatin state segmentation	ENCODE
Weak enhancers (K562)	Chromatin state segmentation	ENCODE
Weak enhancers (NHEK)	Chromatin state segmentation	ENCODE
Weak enhancers (NHLF)	Chromatin state segmentation	ENCODE
Weak promoters (GM12878)	Chromatin state segmentation	ENCODE
Weak promoters (H1-hESC)	Chromatin state segmentation	ENCODE
Weak promoters (HepG2)	Chromatin state segmentation	ENCODE
Weak promoters (HMEC)	Chromatin state segmentation	ENCODE
Weak promoters (HSMM)	Chromatin state segmentation	ENCODE
Weak promoters (HUVEC)	Chromatin state segmentation	ENCODE
Weak promoters (K562)	Chromatin state segmentation	ENCODE
Weak promoters (NHEK)	Chromatin state segmentation	ENCODE
Weak promoters (NHLF)	Chromatin state segmentation	ENCODE
Weak transcribed (GM12878)	Chromatin state segmentation	ENCODE
Weak transcribed (H1-hESC)	Chromatin state segmentation	ENCODE
Weak transcribed (HepG2)	Chromatin state segmentation	ENCODE
Weak transcribed (HMEC)	Chromatin state segmentation	ENCODE
Weak transcribed (HSMM)	Chromatin state segmentation	ENCODE
Weak transcribed (HUVEC)	Chromatin state segmentation	ENCODE
Weak transcribed (K562)	Chromatin state segmentation	ENCODE
Weak transcribed (NHEK)	Chromatin state segmentation	ENCODE
Weak transcribed (NHLF)	Chromatin state segmentation	ENCODE
Conservation, 46-way by Phast-Cons	Conservation	UCSC Genome Browser
CpG islands (specific)	CpG islands	UCSC Genome Browser
HMEC (RRBS)	DNA methylation	ENCODE
HSMM (RRBS)	DNA methylation	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
HepG2 (RRBS)	DNA methylation	ENCODE
H1-hESC (RRBS)	DNA methylation	ENCODE
HeLaS3 (RRBS)	DNA methylation	ENCODE
GM12878 (RRBS)	DNA methylation	ENCODE
DNA	DNA Repeats	UCSC Genome Browser
LINE	DNA Repeats	UCSC Genome Browser
Low complexity	DNA Repeats	UCSC Genome Browser
LTR	DNA Repeats	UCSC Genome Browser
rRNA	DNA Repeats	UCSC Genome Browser
Satellite	DNA Repeats	UCSC Genome Browser
Simple repeats	DNA Repeats	UCSC Genome Browser
SINE	DNA Repeats	UCSC Genome Browser
snRNA	DNA Repeats	UCSC Genome Browser
tRNA	DNA Repeats	UCSC Genome Browser
Unknown	DNA Repeats	UCSC Genome Browser
A frequency	DNA sequence	UCSC Genome Browser
A+T frequency	DNA sequence	UCSC Genome Browser
C frequency	DNA sequence	UCSC Genome Browser
C+G frequency	DNA sequence	UCSC Genome Browser
CpA frequency	DNA sequence	UCSC Genome Browser
CpA+TpG frequency	DNA sequence	UCSC Genome Browser
CpG frequency	DNA sequence	UCSC Genome Browser
G frequency	DNA sequence	UCSC Genome Browser
T frequency	DNA sequence	UCSC Genome Browser
TpG frequency	DNA sequence	UCSC Genome Browser
DNaseI (GM12878)	DNaseI hypersensitive sites	ENCODE
DNaseI (H1-hESC)	DNaseI hypersensitive sites	ENCODE
DNaseI (HeLaS3)	DNaseI hypersensitive sites	ENCODE
DNaseI (HepG2)	DNaseI hypersensitive sites	ENCODE
DNaseI (HMEC)	DNaseI hypersensitive sites	ENCODE
DNaseI (HUVEC)	DNaseI hypersensitive sites	ENCODE
DNaseI (K562)	DNaseI hypersensitive sites	ENCODE
DNaseI (NHEK)	DNaseI hypersensitive sites	ENCODE
DNaseI (NHLF)	DNaseI hypersensitive sites	ENCODE
Gene bodies	Genes and annotations	UCSC Genome Browser and Ensembl
Gene exons	Genes and annotations	UCSC Genome Browser and Ensembl
Gene names and symbols	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-10kb to 2kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-1kb to 1kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-5kb to 1kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene transcription start sites	Genes and annotations	UCSC Genome Browser and Ensembl
GO annotations	Genes and annotations	UCSC Genome Browser and Ensembl
OMIM annotations	Genes and annotations	UCSC Genome Browser and Ensembl
CTCF (GM12878)	Histone modifications	ENCODE
CTCF (H1-hESC)	Histone modifications	ENCODE
CTCF (HepG2)	Histone modifications	ENCODE
Continued on next page		



Annotation name	Annotation group	Annotation source
CTCF (HMEC)	Histone modifications	ENCODE
CTCF (HSMM)	Histone modifications	ENCODE
CTCF (HUVEC)	Histone modifications	ENCODE
CTCF (K562)	Histone modifications	ENCODE
CTCF (NHEK)	Histone modifications	ENCODE
CTCF (NHLF)	Histone modifications	ENCODE
CTCF (Osteobl)	Histone modifications	ENCODE
CTCF (HSMMtube)	Histone modifications	ENCODE
CTCF (NHA)	Histone modifications	ENCODE
CTCF (NHDFAd)	Histone modifications	ENCODE
H2A.Z (HepG2)	Histone modifications	ENCODE
H2A.Z (Osteobl)	Histone modifications	ENCODE
H2A.Z (K562)	Histone modifications	ENCODE
H2A.Z (HSMM)	Histone modifications	ENCODE
H2A.Z (HSMMtube)	Histone modifications	ENCODE
H2A.Z (GM12878)	Histone modifications	ENCODE
H3K27ac (GM12878)	Histone modifications	ENCODE
H3K27ac (HepG2)	Histone modifications	ENCODE
H3K27ac (HMEC)	Histone modifications	ENCODE
H3K27ac (HSMM)	Histone modifications	ENCODE
H3K27ac (HUVEC)	Histone modifications	ENCODE
H3K27ac (K562)	Histone modifications	ENCODE
H3K27ac (NHEK)	Histone modifications	ENCODE
H3K27ac (NHLF)	Histone modifications	ENCODE
H3K27ac (NHDFAd)	Histone modifications	ENCODE
H3K27ac (Osteobl)	Histone modifications	ENCODE
H3K27ac (H1hESC)	Histone modifications	ENCODE
H3K27ac (HSMMtube)	Histone modifications	ENCODE
H3K27ac (NHA)	Histone modifications	ENCODE
H3K27ac (HelaS3)	Histone modifications	ENCODE
H3K27me3 (GM12878)	Histone modifications	ENCODE
H3K27me3 (H1-hESC)	Histone modifications	ENCODE
H3K27me3 (HMEC)	Histone modifications	ENCODE
H3K27me3 (HSMM)	Histone modifications	ENCODE
H3K27me3 (HUVEC)	Histone modifications	ENCODE
H3K27me3 (K562)	Histone modifications	ENCODE
H3K27me3 (NHEK)	Histone modifications	ENCODE
H3K27me3 (NHLF)	Histone modifications	ENCODE
H3K27me3 (HepG2)	Histone modifications	ENCODE
H3K27me3 (NHDFAd)	Histone modifications	ENCODE
H3K27me3 (NHA)	Histone modifications	ENCODE
H3K27me3 (HelaS3)	Histone modifications	ENCODE
H3K36me3 (GM12878)	Histone modifications	ENCODE
H3K36me3 (H1-hESC)	Histone modifications	ENCODE
H3K36me3 (HepG2)	Histone modifications	ENCODE
H3K36me3 (HMEC)	Histone modifications	ENCODE
H3K36me3 (HSMM)	Histone modifications	ENCODE
H3K36me3 (HUVEC)	Histone modifications	ENCODE
H3K36me3 (K562)	Histone modifications	ENCODE
H3K36me3 (NHEK)	Histone modifications	ENCODE
H3K36me3 (NHLF)	Histone modifications	ENCODE
H3K36me3 (NHDFAd)	Histone modifications	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
H3K36me3 (Osteobl)	Histone modifications	ENCODE
H3K36me3 (HSMMtube)	Histone modifications	ENCODE
H3K36me3 (NHA)	Histone modifications	ENCODE
H3K36me3 (HelaS3)	Histone modifications	ENCODE
H3K4me1 (GM12878)	Histone modifications	ENCODE
H3K4me1 (H1-hESC)	Histone modifications	ENCODE
H3K4me1 (HMEC)	Histone modifications	ENCODE
H3K4me1 (HSMM)	Histone modifications	ENCODE
H3K4me1 (HUVEC)	Histone modifications	ENCODE
H3K4me1 (K562)	Histone modifications	ENCODE
H3K4me1 (NHEK)	Histone modifications	ENCODE
H3K4me1 (NHLF)	Histone modifications	ENCODE
H3K4me1 (Osteobl)	Histone modifications	ENCODE
H3K4me1 (HSMMtube)	Histone modifications	ENCODE
H3K4me1 (NHA)	Histone modifications	ENCODE
H3K4me2 (GM12878)	Histone modifications	ENCODE
H3K4me2 (H1-hESC)	Histone modifications	ENCODE
H3K4me2 (HepG2)	Histone modifications	ENCODE
H3K4me2 (HMEC)	Histone modifications	ENCODE
H3K4me2 (HSMM)	Histone modifications	ENCODE
H3K4me2 (HUVEC)	Histone modifications	ENCODE
H3K4me2 (K562)	Histone modifications	ENCODE
H3K4me2 (NHEK)	Histone modifications	ENCODE
H3K4me2 (NHLF)	Histone modifications	ENCODE
H3K4me2 (HSMMtube)	Histone modifications	ENCODE
H3K4me2 (HelaS3)	Histone modifications	ENCODE
H3K4me2 (NHDFAd)	Histone modifications	ENCODE
H3K4me2 (Osteobl)	Histone modifications	ENCODE
H3K4me3 (GM12878)	Histone modifications	ENCODE
H3K4me3 (H1-hESC)	Histone modifications	ENCODE
H3K4me3 (HepG2)	Histone modifications	ENCODE
H3K4me3 (HMEC)	Histone modifications	ENCODE
H3K4me3 (HSMM)	Histone modifications	ENCODE
H3K4me3 (HUVEC)	Histone modifications	ENCODE
H3K4me3 (K562)	Histone modifications	ENCODE
H3K4me3 (NHEK)	Histone modifications	ENCODE
H3K4me3 (NHLF)	Histone modifications	ENCODE
H3K4me3 (HSMMtube)	Histone modifications	ENCODE
H3K4me3 (NHDFAd)	Histone modifications	ENCODE
H3K4me3 (HelaS3)	Histone modifications	ENCODE
H3K4me3 (NHA)	Histone modifications	ENCODE
H3K9ac (GM12878)	Histone modifications	ENCODE
H3K9ac (H1-hESC)	Histone modifications	ENCODE
H3K9ac (HepG2)	Histone modifications	ENCODE
H3K9ac (HMEC)	Histone modifications	ENCODE
H3K9ac (HSMM)	Histone modifications	ENCODE
H3K9ac (HUVEC)	Histone modifications	ENCODE
H3K9ac (K562)	Histone modifications	ENCODE
H3K9ac (NHEK)	Histone modifications	ENCODE
H3K9ac (NHLF)	Histone modifications	ENCODE
H3K9ac (HSMMtube)	Histone modifications	ENCODE
H3K9ac (HelaS3)	Histone modifications	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
H3K9ac (NHDFAd)	Histone modifications	ENCODE
H3K9me1 (HUVEC)	Histone modifications	ENCODE
H3K9me1 (K562)	Histone modifications	ENCODE
H3K9me1 (NHEK)	Histone modifications	ENCODE
H3K9me3 (GM12878)	Histone modifications	ENCODE
H3K9me3 (Osteobl)	Histone modifications	ENCODE
H3K9me3 (HSMM)	Histone modifications	ENCODE
H3K9me3 (K562)	Histone modifications	ENCODE
H3K79me2 (HepG2)	Histone modifications	ENCODE
H3K79me2 (K562)	Histone modifications	ENCODE
H3K79me2 (HSMM)	Histone modifications	ENCODE
H3K79me2 (HSMMtube)	Histone modifications	ENCODE
H3K79me2 (GM12878)	Histone modifications	ENCODE
H3K79me2 (HelaS3)	Histone modifications	ENCODE
H4K20me1 (GM12878)	Histone modifications	ENCODE
H4K20me1 (H1-hESC)	Histone modifications	ENCODE
H4K20me1 (HepG2)	Histone modifications	ENCODE
H4K20me1 (HMEC)	Histone modifications	ENCODE
H4K20me1 (HSMM)	Histone modifications	ENCODE
H4K20me1 (HUVEC)	Histone modifications	ENCODE
H4K20me1 (K562)	Histone modifications	ENCODE
H4K20me1 (NHEK)	Histone modifications	ENCODE
H4K20me1 (NHLF)	Histone modifications	ENCODE
H4K20me1 (HSMMtube)	Histone modifications	ENCODE
H4K20me1 (HelaS3)	Histone modifications	ENCODE
Pol2(b) (HUVEC)	Histone modifications	ENCODE
Pol2(b) (K562)	Histone modifications	ENCODE
Pol2(b) (HelaS3)	Histone modifications	ENCODE
Pol2(b) (NHEK)	Histone modifications	ENCODE
Lamina associated domains	Lamina associated domains	UCSC Genome Browser
MAX (K562)	Transcription factor binding sites	UCSC Genome Browser
PU.1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
PU.1 (K562)	Transcription factor binding sites	UCSC Genome Browser
Sp1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Sp1 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
Sp1 (K562)	Transcription factor binding sites	UCSC Genome Browser
Sp1 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
SRF (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
SRF (K562)	Transcription factor binding sites	UCSC Genome Browser
SRF (GM12878)	Transcription factor binding sites	UCSC Genome Browser
SRF (HepG2)	Transcription factor binding sites	UCSC Genome Browser
YY1 (K562)	Transcription factor binding sites	UCSC Genome Browser
ATF3 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
ATF3 (K562)	Transcription factor binding sites	UCSC Genome Browser
ATF3 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
ATF3 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
BATF (GM12878)	Transcription factor binding sites	UCSC Genome Browser
BCL3 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
EGR-1 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
EGR-1 (K562)	Transcription factor binding sites	UCSC Genome Browser
EGR-1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
ETS1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
ETS1 (K562)	Transcription factor binding sites	UCSC Genome Browser
GABP (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
GABP (K562)	Transcription factor binding sites	UCSC Genome Browser
GABP (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
GABP (GM12878)	Transcription factor binding sites	UCSC Genome Browser
JunD (HepG2)	Transcription factor binding sites	UCSC Genome Browser
JunD (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
NRSF (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
NRSF (K562)	Transcription factor binding sites	UCSC Genome Browser
NRSF (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
NRSF (GM12878)	Transcription factor binding sites	UCSC Genome Browser
NRSF (HepG2)	Transcription factor binding sites	UCSC Genome Browser
P300 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
P300 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
P300 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
PBX3 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (K562)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
RXRa (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
RXRa (HepG2)	Transcription factor binding sites	UCSC Genome Browser
RXRa (GM12878)	Transcription factor binding sites	UCSC Genome Browser
SIX5 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
SIX5 (K562)	Transcription factor binding sites	UCSC Genome Browser
SIX5 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
TAF1 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
TAF1 (K562)	Transcription factor binding sites	UCSC Genome Browser
TAF1 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
TAF1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
TAF1 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
USF1 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
USF1 (K562)	Transcription factor binding sites	UCSC Genome Browser
USF1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
USF1 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
FOSL2 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
MEF2A (K562)	Transcription factor binding sites	UCSC Genome Browser
MEF2A (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (K562)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
TCF12 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
TCF12 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
BCL11A (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
BCL11A (GM12878)	Transcription factor binding sites	UCSC Genome Browser
BCLAF1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
BCLAF1 (K562)	Transcription factor binding sites	UCSC Genome Browser
POU2F2 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
ZBTB33 (K562)	Transcription factor binding sites	UCSC Genome Browser
ZBTB33 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
ZBTB33 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
BHLNE40 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
PAX5-C19 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
PAX5-C20 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Sin3Ak-20 (H1hESC)	Transcription factor binding sites	UCSC Genome Browser
Sin3Ak-20 (K562)	Transcription factor binding sites	UCSC Genome Browser
Sin3Ak-20 (HepG2)	Transcription factor binding sites	UCSC Genome Browser

### A.1.2. hg18

Table A.2.: Full listing of the EpiExplorer's annotation datasets for human genome assembly hg18.

Annotation name	Annotation group	Annotation source
Active promoters (GM12878)	Chromatin state segmentation	ENCODE
Active promoters (H1-hESC)	Chromatin state segmentation	ENCODE
Active promoters (HepG2)	Chromatin state segmentation	ENCODE
Active promoters (HMEC)	Chromatin state segmentation	ENCODE
Active promoters (HSMM)	Chromatin state segmentation	ENCODE
Active promoters (HUVEC)	Chromatin state segmentation	ENCODE
Active promoters (K562)	Chromatin state segmentation	ENCODE
Active promoters (NHEK)	Chromatin state segmentation	ENCODE
Active promoters (NHLF)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (GM12878)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (H1-hESC)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HepG2)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HMEC)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HSMM)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (HUVEC)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (K562)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (NHEK)	Chromatin state segmentation	ENCODE
Heterochromatin (low signal) (NHLF)	Chromatin state segmentation	ENCODE
Insulators (GM12878)	Chromatin state segmentation	ENCODE
Insulators (H1-hESC)	Chromatin state segmentation	ENCODE
Insulators (HepG2)	Chromatin state segmentation	ENCODE
Insulators (HMEC)	Chromatin state segmentation	ENCODE
Insulators (HSMM)	Chromatin state segmentation	ENCODE
Insulators (HUVEC)	Chromatin state segmentation	ENCODE
Insulators (K562)	Chromatin state segmentation	ENCODE
Insulators (NHEK)	Chromatin state segmentation	ENCODE
Insulators (NHLF)	Chromatin state segmentation	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
Poised promoters (GM12878)	Chromatin state segmentation	ENCODE
Poised promoters (H1-hESC)	Chromatin state segmentation	ENCODE
Poised promoters (HepG2)	Chromatin state segmentation	ENCODE
Poised promoters (HMEC)	Chromatin state segmentation	ENCODE
Poised promoters (HSMM)	Chromatin state segmentation	ENCODE
Poised promoters (HUVEC)	Chromatin state segmentation	ENCODE
Poised promoters (K562)	Chromatin state segmentation	ENCODE
Poised promoters (NHEK)	Chromatin state segmentation	ENCODE
Poised promoters (NHLF)	Chromatin state segmentation	ENCODE
Polycomb repressed (GM12878)	Chromatin state segmentation	ENCODE
Polycomb repressed (H1-hESC)	Chromatin state segmentation	ENCODE
Polycomb repressed (HepG2)	Chromatin state segmentation	ENCODE
Polycomb repressed (HMEC)	Chromatin state segmentation	ENCODE
Polycomb repressed (HSMM)	Chromatin state segmentation	ENCODE
Polycomb repressed (HUVEC)	Chromatin state segmentation	ENCODE
Polycomb repressed (K562)	Chromatin state segmentation	ENCODE
Polycomb repressed (NHEK)	Chromatin state segmentation	ENCODE
Polycomb repressed (NHLF)	Chromatin state segmentation	ENCODE
Repetitive CNV (GM12878)	Chromatin state segmentation	ENCODE
Repetitive CNV (H1-hESC)	Chromatin state segmentation	ENCODE
Repetitive CNV (HepG2)	Chromatin state segmentation	ENCODE
Repetitive CNV (HMEC)	Chromatin state segmentation	ENCODE
Repetitive CNV (HSMM)	Chromatin state segmentation	ENCODE
Repetitive CNV (HUVEC)	Chromatin state segmentation	ENCODE
Repetitive CNV (K562)	Chromatin state segmentation	ENCODE
Repetitive CNV (NHEK)	Chromatin state segmentation	ENCODE
Repetitive CNV (NHLF)	Chromatin state segmentation	ENCODE
Strong enhancers (GM12878)	Chromatin state segmentation	ENCODE
Strong enhancers (H1-hESC)	Chromatin state segmentation	ENCODE
Strong enhancers (HepG2)	Chromatin state segmentation	ENCODE
Strong enhancers (HMEC)	Chromatin state segmentation	ENCODE
Strong enhancers (HSMM)	Chromatin state segmentation	ENCODE
Strong enhancers (HUVEC)	Chromatin state segmentation	ENCODE
Strong enhancers (K562)	Chromatin state segmentation	ENCODE
Strong enhancers (NHEK)	Chromatin state segmentation	ENCODE
Strong enhancers (NHLF)	Chromatin state segmentation	ENCODE
Transcriptional elongation (GM12878)	Chromatin state segmentation	ENCODE
Transcriptional elongation (H1-hESC)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HepG2)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HMEC)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HSMM)	Chromatin state segmentation	ENCODE
Transcriptional elongation (HUVEC)	Chromatin state segmentation	ENCODE
Transcriptional elongation (K562)	Chromatin state segmentation	ENCODE
Transcriptional elongation (NHEK)	Chromatin state segmentation	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
Transcriptional elongation (NHLF)	Chromatin state segmentation	ENCODE
Transcriptional transition (GM12878)	Chromatin state segmentation	ENCODE
Transcriptional transition (H1-hESC)	Chromatin state segmentation	ENCODE
Transcriptional transition (HepG2)	Chromatin state segmentation	ENCODE
Transcriptional transition (HMEC)	Chromatin state segmentation	ENCODE
Transcriptional transition (HSMM)	Chromatin state segmentation	ENCODE
Transcriptional transition (HUVEC)	Chromatin state segmentation	ENCODE
Transcriptional transition (K562)	Chromatin state segmentation	ENCODE
Transcriptional transition (NHEK)	Chromatin state segmentation	ENCODE
Transcriptional transition (NHLF)	Chromatin state segmentation	ENCODE
Weak enhancers (GM12878)	Chromatin state segmentation	ENCODE
Weak enhancers (H1-hESC)	Chromatin state segmentation	ENCODE
Weak enhancers (HepG2)	Chromatin state segmentation	ENCODE
Weak enhancers (HMEC)	Chromatin state segmentation	ENCODE
Weak enhancers (HSMM)	Chromatin state segmentation	ENCODE
Weak enhancers (HUVEC)	Chromatin state segmentation	ENCODE
Weak enhancers (K562)	Chromatin state segmentation	ENCODE
Weak enhancers (NHEK)	Chromatin state segmentation	ENCODE
Weak enhancers (NHLF)	Chromatin state segmentation	ENCODE
Weak promoters (GM12878)	Chromatin state segmentation	ENCODE
Weak promoters (H1-hESC)	Chromatin state segmentation	ENCODE
Weak promoters (HepG2)	Chromatin state segmentation	ENCODE
Weak promoters (HMEC)	Chromatin state segmentation	ENCODE
Weak promoters (HSMM)	Chromatin state segmentation	ENCODE
Weak promoters (HUVEC)	Chromatin state segmentation	ENCODE
Weak promoters (K562)	Chromatin state segmentation	ENCODE
Weak promoters (NHEK)	Chromatin state segmentation	ENCODE
Weak promoters (NHLF)	Chromatin state segmentation	ENCODE
Weak transcribed (GM12878)	Chromatin state segmentation	ENCODE
Weak transcribed (H1-hESC)	Chromatin state segmentation	ENCODE
Weak transcribed (HepG2)	Chromatin state segmentation	ENCODE
Weak transcribed (HMEC)	Chromatin state segmentation	ENCODE
Weak transcribed (HSMM)	Chromatin state segmentation	ENCODE
Weak transcribed (HUVEC)	Chromatin state segmentation	ENCODE
Weak transcribed (K562)	Chromatin state segmentation	ENCODE
Weak transcribed (NHEK)	Chromatin state segmentation	ENCODE
Weak transcribed (NHLF)	Chromatin state segmentation	ENCODE
28-way most conserved elements	Conservation	UCSC Genome Browser
CpG islands (sensitive)	CpG islands	CGIHunter
CpG islands (specific)	CpG islands	UCSC Genome Browser
Fetal brain (RRBS)	DNA methylation	ROADMAP
Fetal heart (RRBS)	DNA methylation	ROADMAP
Fetal kidney (RRBS)	DNA methylation	ROADMAP
Continued on next page		



Annotation name	Annotation group	Annotation source
Fetal lung (RRBS)	DNA methylation	ROADMAP
hEB16d H1 p38 (RRBS)	DNA methylation	ROADMAP
hES H1 p38 (RRBS)	DNA methylation	ROADMAP
hES H9 p58 (RRBS)	DNA methylation	ROADMAP
hFib 11 p8 (RRBS)	DNA methylation	ROADMAP
Human blood CD34 mobilized REMC (RRBS)	DNA methylation	ROADMAP
Neuron H9 derived (RRBS)	DNA methylation	ROADMAP
NPC H9 derived (RRBS)	DNA methylation	ROADMAP
Skeletal muscle (RRBS)	DNA methylation	ROADMAP
Smooth muscle (RRBS)	DNA methylation	ROADMAP
Stomach mucosa (RRBS)	DNA methylation	ROADMAP
DNA	DNA Repeats	UCSC Genome Browser
LINE	DNA Repeats	UCSC Genome Browser
Low complexity	DNA Repeats	UCSC Genome Browser
LTR	DNA Repeats	UCSC Genome Browser
rRNA	DNA Repeats	UCSC Genome Browser
Satellite	DNA Repeats	UCSC Genome Browser
Simple repeats	DNA Repeats	UCSC Genome Browser
SINE	DNA Repeats	UCSC Genome Browser
snRNA	DNA Repeats	UCSC Genome Browser
tRNA	DNA Repeats	UCSC Genome Browser
Unknown	DNA Repeats	UCSC Genome Browser
A frequency	DNA sequence	UCSC Genome Browser
A+T frequency	DNA sequence	UCSC Genome Browser
C frequency	DNA sequence	UCSC Genome Browser
C+G frequency	DNA sequence	UCSC Genome Browser
CpA frequency	DNA sequence	UCSC Genome Browser
CpA+TpG frequency	DNA sequence	UCSC Genome Browser
CpG frequency	DNA sequence	UCSC Genome Browser
G frequency	DNA sequence	UCSC Genome Browser
T frequency	DNA sequence	UCSC Genome Browser
TpG frequency	DNA sequence	UCSC Genome Browser
DNaseI (GM12878)	DNaseI hypersensitive sites	ENCODE
DNaseI (H1-hESC)	DNaseI hypersensitive sites	ENCODE
DNaseI (HelaS3)	DNaseI hypersensitive sites	ENCODE
DNaseI (HepG2)	DNaseI hypersensitive sites	ENCODE
DNaseI (HMEC)	DNaseI hypersensitive sites	ENCODE
DNaseI (HUVEC)	DNaseI hypersensitive sites	ENCODE
DNaseI (K562)	DNaseI hypersensitive sites	ENCODE
DNaseI (NHEK)	DNaseI hypersensitive sites	ENCODE
DNaseI (NHLF)	DNaseI hypersensitive sites	ENCODE
Gene bodies	Genes and annotations	UCSC Genome Browser and Ensembl
Gene exons	Genes and annotations	UCSC Genome Browser and Ensembl
Gene names and symbols	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-10kb to 2kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-1kb to 1kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-5kb to 1kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Continued on next page		

Annotation name	Annotation group	Annotation source
Gene transcription start sites	Genes and annotations	UCSC Genome Browser and Ensembl
GO annotations	Genes and annotations	UCSC Genome Browser and Ensembl
OMIM annotations	Genes and annotations	UCSC Genome Browser and Ensembl
CTCF (GM12878)	Histone modifications	ENCODE
CTCF (H1-hESC)	Histone modifications	ENCODE
CTCF (HepG2)	Histone modifications	ENCODE
CTCF (HMEC)	Histone modifications	ENCODE
CTCF (HSMM)	Histone modifications	ENCODE
CTCF (HUVEC)	Histone modifications	ENCODE
CTCF (K562)	Histone modifications	ENCODE
CTCF (NHEK)	Histone modifications	ENCODE
CTCF (NHLF)	Histone modifications	ENCODE
H3K27ac (GM12878)	Histone modifications	ENCODE
H3K27ac (HepG2)	Histone modifications	ENCODE
H3K27ac (HMEC)	Histone modifications	ENCODE
H3K27ac (HSMM)	Histone modifications	ENCODE
H3K27ac (HUVEC)	Histone modifications	ENCODE
H3K27ac (K562)	Histone modifications	ENCODE
H3K27ac (NHEK)	Histone modifications	ENCODE
H3K27ac (NHLF)	Histone modifications	ENCODE
H3K27me3 (GM12878)	Histone modifications	ENCODE
H3K27me3 (H1-hESC)	Histone modifications	ENCODE
H3K27me3 (HMEC)	Histone modifications	ENCODE
H3K27me3 (HSMM)	Histone modifications	ENCODE
H3K27me3 (HUVEC)	Histone modifications	ENCODE
H3K27me3 (K562)	Histone modifications	ENCODE
H3K27me3 (NHEK)	Histone modifications	ENCODE
H3K27me3 (NHLF)	Histone modifications	ENCODE
H3K36me3 (GM12878)	Histone modifications	ENCODE
H3K36me3 (H1-hESC)	Histone modifications	ENCODE
H3K36me3 (HepG2)	Histone modifications	ENCODE
H3K36me3 (HMEC)	Histone modifications	ENCODE
H3K36me3 (HSMM)	Histone modifications	ENCODE
H3K36me3 (HUVEC)	Histone modifications	ENCODE
H3K36me3 (K562)	Histone modifications	ENCODE
H3K36me3 (NHEK)	Histone modifications	ENCODE
H3K36me3 (NHLF)	Histone modifications	ENCODE
H3K4me1 (GM12878)	Histone modifications	ENCODE
H3K4me1 (H1-hESC)	Histone modifications	ENCODE
H3K4me1 (HMEC)	Histone modifications	ENCODE
H3K4me1 (HSMM)	Histone modifications	ENCODE
H3K4me1 (HUVEC)	Histone modifications	ENCODE
H3K4me1 (K562)	Histone modifications	ENCODE
H3K4me1 (NHEK)	Histone modifications	ENCODE
H3K4me1 (NHLF)	Histone modifications	ENCODE
H3K4me2 (GM12878)	Histone modifications	ENCODE
H3K4me2 (H1-hESC)	Histone modifications	ENCODE
H3K4me2 (HepG2)	Histone modifications	ENCODE
H3K4me2 (HMEC)	Histone modifications	ENCODE
H3K4me2 (HSMM)	Histone modifications	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
H3K4me2 (HUVEC)	Histone modifications	ENCODE
H3K4me2 (K562)	Histone modifications	ENCODE
H3K4me2 (NHEK)	Histone modifications	ENCODE
H3K4me2 (NHLF)	Histone modifications	ENCODE
H3K4me3 (GM12878)	Histone modifications	ENCODE
H3K4me3 (H1-hESC)	Histone modifications	ENCODE
H3K4me3 (HepG2)	Histone modifications	ENCODE
H3K4me3 (HMEC)	Histone modifications	ENCODE
H3K4me3 (HSMM)	Histone modifications	ENCODE
H3K4me3 (HUVEC)	Histone modifications	ENCODE
H3K4me3 (K562)	Histone modifications	ENCODE
H3K4me3 (NHEK)	Histone modifications	ENCODE
H3K4me3 (NHLF)	Histone modifications	ENCODE
H3K9ac (GM12878)	Histone modifications	ENCODE
H3K9ac (H1-hESC)	Histone modifications	ENCODE
H3K9ac (HepG2)	Histone modifications	ENCODE
H3K9ac (HMEC)	Histone modifications	ENCODE
H3K9ac (HSMM)	Histone modifications	ENCODE
H3K9ac (HUVEC)	Histone modifications	ENCODE
H3K9ac (K562)	Histone modifications	ENCODE
H3K9ac (NHEK)	Histone modifications	ENCODE
H3K9ac (NHLF)	Histone modifications	ENCODE
H3K9me1 (HUVEC)	Histone modifications	ENCODE
H3K9me1 (K562)	Histone modifications	ENCODE
H3K9me1 (NHEK)	Histone modifications	ENCODE
H4K20me1 (GM12878)	Histone modifications	ENCODE
H4K20me1 (H1-hESC)	Histone modifications	ENCODE
H4K20me1 (HepG2)	Histone modifications	ENCODE
H4K20me1 (HMEC)	Histone modifications	ENCODE
H4K20me1 (HSMM)	Histone modifications	ENCODE
H4K20me1 (HUVEC)	Histone modifications	ENCODE
H4K20me1 (K562)	Histone modifications	ENCODE
H4K20me1 (NHEK)	Histone modifications	ENCODE
H4K20me1 (NHLF)	Histone modifications	ENCODE
Pol2(b) (HUVEC)	Histone modifications	ENCODE
Pol2(b) (K562)	Histone modifications	ENCODE
Pol2(b) (NHEK)	Histone modifications	ENCODE
Lamina associated domains	Lamina associated domains	UCSC Genome Browser
AP2alpha (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
AP2gamma (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
ATF3 (K562)	Transcription factor binding sites	UCSC Genome Browser
BDP1 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
BDP1 (K562)	Transcription factor binding sites	UCSC Genome Browser
BRF1 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
BRF1 (K562)	Transcription factor binding sites	UCSC Genome Browser
BRF2 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
BRF2 (K562)	Transcription factor binding sites	UCSC Genome Browser
cFos (GM12878)	Transcription factor binding sites	UCSC Genome Browser
cFos (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
cFos (K562)	Transcription factor binding sites	UCSC Genome Browser
cJun (GM12878)	Transcription factor binding sites	UCSC Genome Browser
cJun (HUVEC)	Transcription factor binding sites	UCSC Genome Browser
cJun (K562)	Transcription factor binding sites	UCSC Genome Browser
cMyc (GM12878)	Transcription factor binding sites	UCSC Genome Browser
cMyc (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
cMyc (K562)	Transcription factor binding sites	UCSC Genome Browser
E2F1 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
E2F4 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
E2F4 (K562)	Transcription factor binding sites	UCSC Genome Browser
E2F6 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
E2F6 (K562)	Transcription factor binding sites	UCSC Genome Browser
GATA1 (K562)	Transcription factor binding sites	UCSC Genome Browser
GATA2 (K562)	Transcription factor binding sites	UCSC Genome Browser
GTF2B (K562)	Transcription factor binding sites	UCSC Genome Browser
HELFc (K562)	Transcription factor binding sites	UCSC Genome Browser
junD (GM12878)	Transcription factor binding sites	UCSC Genome Browser
junD (K562)	Transcription factor binding sites	UCSC Genome Browser
MAX (GM12878)	Transcription factor binding sites	UCSC Genome Browser
MAX (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
MAX (HUVEC)	Transcription factor binding sites	UCSC Genome Browser
MAX (K562)	Transcription factor binding sites	UCSC Genome Browser
NFE2 (K562)	Transcription factor binding sites	UCSC Genome Browser
NFKB (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (HUVEC)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (K562)	Transcription factor binding sites	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
Pol3 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
Pol3 (K562)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (K562)	Transcription factor binding sites	UCSC Genome Browser
RPC155 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
RPC155 (K562)	Transcription factor binding sites	UCSC Genome Browser
SETDB1 (K562)	Transcription factor binding sites	UCSC Genome Browser
SIRT6 (K562)	Transcription factor binding sites	UCSC Genome Browser
TFIIIC110 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
TFIIIC110 (K562)	Transcription factor binding sites	UCSC Genome Browser
TR4 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
TR4 (HeLaS3)	Transcription factor binding sites	UCSC Genome Browser
TR4 (HepG2)	Transcription factor binding sites	UCSC Genome Browser
TR4 (K562)	Transcription factor binding sites	UCSC Genome Browser
XRCC4 (K562)	Transcription factor binding sites	UCSC Genome Browser
YY1 (GM12878)	Transcription factor binding sites	UCSC Genome Browser
YY1 (K562)	Transcription factor binding sites	UCSC Genome Browser
ZNF263 (K562)	Transcription factor binding sites	UCSC Genome Browser
ZNF274 (K562)	Transcription factor binding sites	UCSC Genome Browser
ZZZ3 (K562)	Transcription factor binding sites	UCSC Genome Browser

## A.2. Mouse genome

### A.2.1. mm9

Table A.3.: Full listing of the EpiExplorer's annotation datasets for mouse genome assembly mm9.

Annotation name	Annotation group	Annotation source
30-way most conserved elements	Conservation	UCSC Genome Browser
CpG islands (sensitive)	CpG islands	CGIHunter
CpG islands (specific)	CpG islands	UCSC Genome Browser
Brain (RRBS)	DNA methylation	ROADMAP
Heart (RRBS)	DNA methylation	ROADMAP
Liver (RRBS)	DNA methylation	ROADMAP
DNA	DNA Repeats	UCSC Genome Browser
LINE	DNA Repeats	UCSC Genome Browser
Low complexity	DNA Repeats	UCSC Genome Browser
LTR	DNA Repeats	UCSC Genome Browser
rRNA	DNA Repeats	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
Satellite	DNA Repeats	UCSC Genome Browser
Simple repeats	DNA Repeats	UCSC Genome Browser
SINE	DNA Repeats	UCSC Genome Browser
snRNA	DNA Repeats	UCSC Genome Browser
tRNA	DNA Repeats	UCSC Genome Browser
Unknown	DNA Repeats	UCSC Genome Browser
A frequency	DNA sequence	UCSC Genome Browser
A+T frequency	DNA sequence	UCSC Genome Browser
C frequency	DNA sequence	UCSC Genome Browser
C+G frequency	DNA sequence	UCSC Genome Browser
CpA frequency	DNA sequence	UCSC Genome Browser
CpA+TpG frequency	DNA sequence	UCSC Genome Browser
CpG frequency	DNA sequence	UCSC Genome Browser
G frequency	DNA sequence	UCSC Genome Browser
T frequency	DNA sequence	UCSC Genome Browser
TpG frequency	DNA sequence	UCSC Genome Browser
DNaseI adult(A20)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Bcellcd19p)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Bcellcd43n)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Cerebellum)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Cerebrum)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Fat)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Fibroblast)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Kidney)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Liver)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Lung)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Tnaive)	DNaseI hypersensitive sites	ENCODE
DNaseI adult(Wholebrain)	DNaseI hypersensitive sites	ENCODE
DNaseI E0 (Escj7S129)	DNaseI hypersensitive sites	ENCODE
DNaseI E0 (Zhbt4129ola)	DNaseI hypersensitive sites	ENCODE
DNaseI E14.5 (Brain)	DNaseI hypersensitive sites	ENCODE
DNaseI immortal (3134Riii)	DNaseI hypersensitive sites	ENCODE
DNaseI immortal (PatskiSpbl6)	DNaseI hypersensitive sites	ENCODE
Gene bodies	Genes and annotations	UCSC Genome Browser and Ensembl
Gene exons	Genes and annotations	UCSC Genome Browser and Ensembl
Gene names and symbols	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-10kb to 2kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-1kb to 1kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene promoters (-5kb to 1kb)	Genes and annotations	UCSC Genome Browser and Ensembl
Gene transcription start sites	Genes and annotations	UCSC Genome Browser and Ensembl
GO annotations	Genes and annotations	UCSC Genome Browser and Ensembl
H3K4me1 (Bmarrow)	Histone modifications	ENCODE
H3K4me1 (Cbellum)	Histone modifications	ENCODE
H3K4me1 (Cortex)	Histone modifications	ENCODE
H3K4me1 (Heart)	Histone modifications	ENCODE
H3K4me1 (Kidney)	Histone modifications	ENCODE
Continued on next page		

Annotation name	Annotation group	Annotation source
H3K4me1 (Liver)	Histone modifications	ENCODE
H3K4me1 (Lung)	Histone modifications	ENCODE
H3K4me1 (Mef)	Histone modifications	ENCODE
H3K4me1 (Spleen)	Histone modifications	ENCODE
H3K4me3 (Bmarrow)	Histone modifications	ENCODE
H3K4me3 (Cbellum)	Histone modifications	ENCODE
H3K4me3 (Cortex)	Histone modifications	ENCODE
H3K4me3 (Heart)	Histone modifications	ENCODE
H3K4me3 (Kidney)	Histone modifications	ENCODE
H3K4me3 (Liver)	Histone modifications	ENCODE
H3K4me3 (Lung)	Histone modifications	ENCODE
H3K4me3 (Mef)	Histone modifications	ENCODE
H3K4me3 (Spleen)	Histone modifications	ENCODE
Lamina associated domains	Lamina associated domains	UCSC Genome Browser
MAX (MEL)	Transcription factor binding sites	UCSC Genome Browser
MAX (CH12)	Transcription factor binding sites	UCSC Genome Browser
TBP (MEL)	Transcription factor binding sites	UCSC Genome Browser
TBP (CH12)	Transcription factor binding sites	UCSC Genome Browser
CHD2 (MEL)	Transcription factor binding sites	UCSC Genome Browser
CHD2 (CH12)	Transcription factor binding sites	UCSC Genome Browser
cJun (CH12)	Transcription factor binding sites	UCSC Genome Browser
cMyb (MEL)	Transcription factor binding sites	UCSC Genome Browser
cMyc (MEL)	Transcription factor binding sites	UCSC Genome Browser
cMyc (CH12)	Transcription factor binding sites	UCSC Genome Browser
CTCF (MEL)	Transcription factor binding sites	UCSC Genome Browser
CTCF (CH12)	Transcription factor binding sites	UCSC Genome Browser
E2F4 (MEL)	Transcription factor binding sites	UCSC Genome Browser
E2F4 (CH12)	Transcription factor binding sites	UCSC Genome Browser
JunD (MEL)	Transcription factor binding sites	UCSC Genome Browser
JunD (CH12)	Transcription factor binding sites	UCSC Genome Browser
MafK (MEL)	Transcription factor binding sites	UCSC Genome Browser
MafK (CH12)	Transcription factor binding sites	UCSC Genome Browser
MxiI (MEL)	Transcription factor binding sites	UCSC Genome Browser
MxiI (CH12)	Transcription factor binding sites	UCSC Genome Browser
P300 (MEL)	Transcription factor binding sites	UCSC Genome Browser
Pol2 (MEL)	Transcription factor binding sites	UCSC Genome Browser
Continued on next page		

Annotation name	Annotation group	Annotation source
Pol2 (CH12)	Transcription factor binding sites	UCSC Genome Browser
SMC3 (MEL)	Transcription factor binding sites	UCSC Genome Browser
SMC3 (CH12)	Transcription factor binding sites	UCSC Genome Browser
USF2 (MEL)	Transcription factor binding sites	UCSC Genome Browser
USF2 (CH12)	Transcription factor binding sites	UCSC Genome Browser
NELFe (MEL)	Transcription factor binding sites	UCSC Genome Browser
NELFe (CH12)	Transcription factor binding sites	UCSC Genome Browser
GATA1 (MEL)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (MEL)	Transcription factor binding sites	UCSC Genome Browser
Rad21 (CH12)	Transcription factor binding sites	UCSC Genome Browser
Bhlhe40nb100 (CH12)	Transcription factor binding sites	UCSC Genome Browser



## B. EpiGRAPH attribute reference sheet

### B.1. Overview

All attributes calculated by EpiGRAPH are given names according to the following hierarchical naming schema:

$$\langle full\text{-}attribute\text{-}identifier \rangle ::= [\langle window \rangle .] \langle attribute\text{-}group\text{-}name \rangle . \langle attribute\text{-}name \rangle . \langle column\text{-}name \rangle$$

- At the top level ( $\langle attribute\text{-}group\text{-}name \rangle$ ), an attribute group pools a set of biologically related attributes, e.g. DNA sequence patterns in the attribute group "*DNA\_Sequence*" or gene-related attributes in the attribute group "*Genes*".
- The intermediate level ( $\langle attribute\text{-}name \rangle$ ) refers to a set of attributes and columns that are derived from the same dataset, e.g. "*RefSeq\_Genes*" and "*CCDS*", both belonging to the attribute group "*Genes*".
- The bottom level ( $\langle column\text{-}name \rangle$ ) refers to a specific column in the table of attributes that EpiGRAPH calculates. All column names are given based on rules identifying a certain mode of calculation (e.g. frequency of overlap or average score), which are described in more detail below.
- An optional top level ( $\langle window \rangle$ ) is used when the attribute calculation includes not only the genomic regions provided by the input dataset, but also adjacent windows upstream and downstream.

Because the full attribute names are often quite long, we also use a shorthand, which is for example used in the column header of EpiGRAPH's dataset of calculated attributes (downloadable from the EpiGRAPH website via the "*Download Data Table*" button on EpiGRAPH's results overview page). A complete mapping from long to short attribute names is provided in the corresponding *X-GRAP* file, which can be downloaded via the "*Download XML Documentation*" button on the results overview page.

Below, we describe four types of attributes that are calculated in different ways as indicated by the  $\langle column\text{-}name \rangle$  segment of their full attribute name. *DNA sequence* attributes and *patch* attributes are most common within EpiGRAPH.

### B.2. DNA sequence attributes (calculated from pattern frequencies)

The columns of DNA sequence attributes represent the frequency of appearance of a specific DNA sequence pattern (e.g. "*CGCG*") within the genome sequence of the genomic region

specified in the input file. The names of the corresponding columns are composed according to the following rule (" $+$ " stands for string concatenation):

*"Pat\_" + <pattern-string> + "\_freq"*

For every sequence pattern, EpiGRAPH provides the option to compute additional statistics on the pattern occurrence throughout the region, such as standard deviation, skewness and kurtosis of the frequency values. These statistics are computed by dividing the genomic region into subregions and calculating the frequency of the pattern in each subregion. This results in a set of pattern frequencies from which standard deviation, skewness and kurtosis are computed. The names of the corresponding columns are composed according to the following rules:

*"Pat" + <pattern-string> + "\_std"* for standard deviation

*"Pat" + <pattern-string> + "\_skew"* for skewness

*"Pat" + <pattern-string> + "\_kurt"* for kurtosis

All pattern frequencies can be computed either in a strand-specific or non-strand-specific way (strand specificity refers to the genomic plus-strand, not to the direction of transcription of the nearest gene). Strand specificity is indicated in the column names according to the following rule:

*<strand-specific-pattern-string> ::= "plus"/"minus" + <pattern-string>*

### B.3. DNA structure attributes (calculated from oligomers with known structure)

DNA structure predictions are calculated for a given genomic region by sliding a window of fixed size over the region and comparing the DNA sequence pattern in this window with a set of oligomers with known structure (which is described by numerical score values). For example, the predicted helix structure of all possible octamers has been quantified by a set of six numeric scores: *twist*, *roll*, *tilt*, *rise*, *slide* and *shift* ((Gardiner et al., 2003)).

For each score, a new column named *<score-name>* is added, its value being the mean of the scores corresponding to all oligomer hits observed while shifting the sliding window over the genomic region. Similar to the pattern frequency attributes described above, we also report *standard deviation* (*<score-name> + "\_std"*), *skewness* (*<score-name> + "\_skew"*) and *kurtosis* (*<score-name> + "\_kurt"*).

### B.4. Patch attributes (quantifying overlap with sets of genomic regions)

Patch attributes describe the frequency of overlap between the genomic regions in the input dataset and various types of other genome annotations that take the form of genomic regions (e.g. CpG islands, repetitive regions and SNPs). For every patch attribute, three basal columns are introduced, which report general statistics about the overlap between the regions in the input dataset and the patch attribute:

*"overlapRegionsCount"* total number of patch attribute regions overlapping the input region, standardized to 1kb

*"overlapTotalLength"* total length of patch attribute regions overlapping the input region, standardized to 1kb

*"overlapAverageSize"* average size of the overlapping regions

In addition to its three basic features (chromosome, start position, end position), a patch attribute may also contain additional columns that can be numeric (referred to as *score* attributes), binary (*class* attributes) or categorical (*category* attributes), giving rise to additional columns during the attribute calculation.

*Score* attributes give rise to columns with the same name as the score column in the patch attribute and are calculated as weighted averages of the patch regions overlapping with the region specified in the input dataset. Weighting is performed according to the length of overlap.

*Class* attributes give rise to columns with the same name as the class column in the patch attribute and are calculated as the dominant class among the patch regions overlapping with the region specified in the input dataset. Furthermore, distribution statistics for each class are reported in additional columns:

*<class-name>+\_ "+<class-value>+\_overlapRegionsCount"* total number of patch attribute regions with value *<class-value>* for class *<class-name>* overlapping with the input region, standardized to 1kb

*<class-name>+\_ "+<class-value>+\_overlapTotalLength"* total length of patch attribute regions with value *<class-value>* for class *<class-name>* overlapping the input region, standardized to 1kb

*<class-name>+\_ "+<class-value>+\_overlapAverageSize"* average size of the overlapping patch attribute regions with value *<class-value>* for class *<class-name>*

*Category* attributes split the patch attribute into several sub-attributes, and the standard measures of overlap are calculated separately for each category, giving rise to the following columns:

*<category-name>+\_ "+<category-value>+\_overlapRegionsCount"* total number of patch attribute regions with value *<category-value>* for class *<category-name>* overlapping with the input region, standardized to 1kb

*<category-name>+\_ "+<category-value>+\_overlapTotalLength"* total length of patch attribute regions with value *<category-value>* for class *<category-name>* overlapping the input region, standardized to 1kb

*<category-name>+\_ "+<category-value>+\_overlapAverageSize"* average size of the overlapping patch attribute regions with value *<category-value>* for class *<category-name>*

Furthermore, for each category EpiGRAPH reports separate averages for all score attributes. The names of the corresponding columns are composed according to the following rule:

"c"<category-name>+<category-value>+"\_o"<score-name> mean score of all patch attribute regions belonging to the category <category-value> and overlapping with the input region

Genomic strand columns are treated as special type of categories, and the columns derived from a strand-specific patch attribute are named by the same rules, except for the prefix "c" being changed to "s":

"s"<strand-name>+<strand-value>+"\_o"<score-name>

## B.5. Gene attributes (quantifying overlap with genes and exons)

Gene attributes are a special case of patch attributes that take the specific structure of eukaryotic genes (exons and introns) into account. They contain a number of additional columns, with names composed according to the following rules

<attribute-name> + *\_elen* : total length of exonic DNA within the region, standardized to 1kb

<attribute-name> + *\_eno* : total number of exons within the region, standardized to 1kb

<attribute-name> + *\_eavg* : average length of the exons overlapping the target region

<attribute-name> + *\_estd* : standard deviation of the lengths of the exons overlapping the region

<attribute-name> + *\_glen* : total length of genic DNA within the region, standardized to 1kb

<attribute-name> + *\_gno* : total number of genes within the region, standardized to 1kb

<attribute-name> + *\_gavg* : average full of the genes overlapping the region

<attribute-name> + *\_gstd* : standard deviation of the lengths of the genes overlapping the region

<attribute-name> + *\_gcav* : average number of exons per gene

<attribute-name> + *\_gcsd* : standard deviation of the exon number per gene

An up-to-date version of this document can be found here : [http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH\\_Attribute\\_Reference\\_Sheet.pdf](http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH_Attribute_Reference_Sheet.pdf)

# Bibliography

- Abeyasinghe, S. S., Chuzhanova, N., Krawczak, M., Ball, E. V., and Cooper, D. N. (2003). Translocation and gross deletion breakpoints in human inherited disease and cancer I: Nucleotide composition and recombination-associated motifs. *Human mutation*, 22(3):229–244.
- Adams, D., Altucci, L., Antonarakis, S., Ballesteros, J., Beck, S., Bird, A., Bock, C., Boehm, B., Campo, E., Caricasole, A., Dahl, F., Dermitzakis, E., Enver, T., Esteller, M., Estivill, X., Ferguson-Smith, A., Fitzgibbon, J., Flicek, P., Giehl, C., Graf, T., Grosveld, F., Guigo, R., Gut, I., Helin, K., Jarvius, J., Kupperts, R., Lehrach, H., Lengauer, T., Lernmark, A., and Leslie, D. (2012). BLUEPRINT to decode the epigenetic signature written in blood. *Nat Biotechnol*, 30:224–226.
- Allen, E., Horvath, S., Tong, F., Kraft, P., Spiteri, E., Riggs, A., and Marahrens, Y. (2003). High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc Natl Acad Sci USA*, 100:9940–9945.
- Apache Lucene (1999). Apache Lucene. <http://lucene.apache.org>.
- Apache Solr (2004). Apache Solr. <http://lucene.apache.org/solr/>.
- Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25:25–29.
- Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern information retrieval*, volume 463. ACM press New York.
- Bajic, V. B., Tan, S. L., Christoffels, A., Schönbach, C., Lipovich, L., Yang, L., Hofmann, O., Kruger, A., Hide, W., Kai, C., Kawai, J., Hume, D. A., Carninci, P., and Hayashizaki, Y. (2006). Mice and Men: Their Promoter Properties. *PLoS Genet*, 2(4):e54+.
- Bannister, A. J. and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395.
- Barski, A., Cuddapah, S., Cui, K., Roh, T., Schones, D., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837.
- Bast, H. and Weber, I. (2006). Type less, find more: fast autocompletion search with a succinct index. In *Proceedings of the 29th annual international ACM SIGIR conference*

- on Research and development in information retrieval*, SIGIR '06, page 364–371, New York, NY, USA. ACM.
- Bast, H. and Weber, I. (2007). The CompleteSearch engine: Interactive, efficient, and towards IR& DB integration. In *CIDR*, pages 88–95. [www.crdrrdb.org](http://www.crdrrdb.org).
- Beisel, C. and Paro, R. (2011). Silencing chromatin: comparing modes and mechanisms. *Nature Reviews Genetics*, 12(2):123–135.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.
- Berger, S. L., Kouzarides, T., Shiekhata, R., and Shilatifard, A. (2009). An operational definition of epigenetics. *Genes & development*, 23(7):781–783.
- Bernstein, B., Stamatoyannopoulos, J., Costello, J., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M., Beaudet, A., Ecker, J., Farnham, P., Hirst, M., Lander, E., Mikkelsen, T., and Thomson, J. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol*, 28:1045–1048.
- Bernstein, B. E., Meissner, A., and Lander, E. S. (2007). The mammalian epigenome. *Cell*, 128(4):669–681.
- Berry, C., Hannenhalli, S., Leipzig, J., and Bushman, F. (2006). Selection of target sites for mobile DNA integration in the human genome. *PLoS Comput Biol*, 2:e157.
- Bibikova, M., Le, J., Barnes, B., Saedinia-Melnyk, S., Zhou, L., Shen, R., and Gunderson, K. L. (2009). Genome-wide DNA methylation profiling using infinium® assay.
- Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development*, 16(1):6–21.
- Bird, A. (2007). Perceptions of epigenetics. *Nature*, 447(7143):396–398.
- Bird, A. P. (1985). CpG-rich islands and the function of DNA methylation. *Nature*, 321(6067):209–213.
- Birney, E., Andrews, T. D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cuff, J., Curwen, V., Cutts, T., Down, T., Eyra, E., Fernandez-Suarez, X. M., Gane, P., Gibbins, B., Gilbert, J., Hammond, M., Hotz, H., Iyer, V., Jekosch, K., Kahari, A., Kasprzyk, A., Keefe, D., Keenan, S., Lehvaslaiho, H., McVicker, G., Melsopp, C., Meidl, P., Mongin, E., Pettett, R., Potter, S., Proctor, G., Rae, M., Searle, S., Slater, G., Smedley, D., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Storey, R., Ureta-Vidal, A., Woodward, K. C., Cameron, G., Durbin, R., Cox, A., Hubbard, T., and Clamp, M. (2004). An overview of ensembl. *Genome Research*, 14(5):925–928.
- Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816.

- Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a web-based genome analysis tool for experimentalists. *Current Protocols in Molecular Biology* / Edited by Frederick M. Ausubel ... [et Al.], Chapter 19:Unit 19.10.1–21.
- Bock, C. (2008). *Computational Epigenetics*. Doctoral dissertation, Universität des Saarlandes, Saarbrücken.
- Bock, C. (2009). EpiGRAPH Background Information and Supplementary Website. <http://epigraph.mpi-inf.mpg.de/WebGRAPH/faces/Background.html>.
- Bock, C. (2012). Analysing and interpreting DNA methylation data. *Nature Reviews Genetics*, 13(10):705–719.
- Bock, C., Halachev, K., Büch, J., and Lengauer, T. (2009). EpiGRAPH: User-friendly software for statistical analysis and prediction of (epi-) genomic data. *Genome Biol*, 10:R14.
- Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., Ziller, M., Croft, G. F., Amoroso, M. W., and Oakley, D. H. (2011). Reference maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. *Cell*, 144(3):439–452.
- Bock, C. and Lengauer, T. (2008). Computational epigenetics. *Bioinformatics*, 24:1–10.
- Bock, C., Paulsen, M., Tierling, S., Mikeska, T., Lengauer, T., and Walter, J. (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS genetics*, 2(3):e26.
- Bock, C., Tomazou, E. M., Brinkman, A. B., Müller, F., Simmer, F., Gu, H., Jäger, N., Gnirke, A., Stunnenberg, H. G., and Meissner, A. (2010a). Quantitative comparison of genome-wide DNA methylation mapping technologies. *Nature biotechnology*, 28(10):1106–1114.
- Bock, C., Von Kuster, G., Halachev, K., Taylor, J., Nekrutenko, A., and Lengauer, T. (2010b). Web-based analysis of (epi-) genome data using epigraph and galaxy. In *Genetic Variation*, pages 275–296. Springer.
- Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2007). CpG island mapping by epigenome prediction. *PLoS computational biology*, 3(6):e110.
- Bock, C., Walter, J., Paulsen, M., and Lengauer, T. (2008). Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res*, 36:e55.
- Brinkman, A. B., Simmer, F., Ma, K., Kaan, A., Zhu, J., and Stunnenberg, H. G. (2010). Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, 52(3):232–236.

- Brunner, A. L., Johnson, D. S., Kim, S. W. W., Valouev, A., Reddy, T. E., Neff, N. F., Anton, E., Medina, C., Nguyen, L., Chiao, E., Oyolu, C. B., Schroth, G. P., Absher, D. M., Baker, J. C., and Myers, R. M. (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome research*, 19(6):1044–1056.
- Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., Semple, C., Taylor, M., Engstrom, P., Frith, M., Forrest, A., Alkema, W., Tan, S., Plessy, C., Kodzius, R., Ravasi, T., Kasukawa, T., Fukuda, S., Kanamori-Katayama, M., Kitazume, Y., Kawaji, H., Kai, C., Nakamura, M., Konno, H., Nakano, K., Mottagui-Tabar, S., Arner, P., Chesi, A., Gustincich, S., and Persichetti, F. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet*, 38:626–635.
- Cedar, H. and Bergman, Y. (2012). Programming of DNA methylation patterns. *Annual review of biochemistry*, 81:97–117.
- CgiHunter (2010-2013). CgiHunter. <http://cgihunter.bioinf.mpi-inf.mpg.de/>.
- Chen, T., Ueda, Y., Dodge, J. E., Wang, Z., and Li, E. (2003). Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by dnmt3a and dnmt3b. *Molecular and cellular biology*, 23(16):5594–5605.
- Cohen, N., Kenigsberg, E., and Tanay, A. (2011). Primate CpG islands are maintained by heterogeneous evolutionary regimes involving minimal selection. *Cell*, 145:773–786.
- Collas, P. and Dahl, J. A. (2008). Chop it, chip it, check it: the current status of chromatin immunoprecipitation. *Front Biosci*, 13(17):929–943.
- Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, 306:636–640.
- Costantini, M., Clay, O., Auletta, F., and Bernardi, G. (2006). An isochore map of human chromosomes. *Genome Res*, 16:536–541.
- Das, R., Dimitrova, N., Xuan, Z., Rollins, R. A., Haghghi, F., Edwards, J. R., Ju, J., Bestor, T. H., and Zhang, M. Q. (2006). Computational prediction of methylation status in human genomic sequences. *Proceedings of the National Academy of Sciences*, 103(28):10713–10716.
- De, S., Shaknovich, R., Riester, M., Elemento, O., Geng, H., Kormaksson, M., Jiang, Y., Woolcock, B., Johnson, N., Polo, J. M., Cerchietti, L., Gascoyne, R. D., Melnick, A., and Michor, F. (2013). Aberration in DNA Methylation in B-Cell Lymphomas Has a Complex Origin and Increases with Disease Severity. *PLoS Genet*, 9(1):e1003137+.
- De Carvalho, D., You, J., and Jones, P. (2010). DNA methylation and cellular reprogramming. *Trends Cell Biol*, 20:609–617.
- Deaton, A. and Bird, A. (2011). CpG islands and the regulation of transcription. *Genes Dev*, 25:1010–1022.



- Delhommeau, F., Dupont, S., Della Valle, V., James, C., Trannoy, S., Masse, A., Kosmider, O., Le Couedic, J., Robert, F., Alberdi, A., Lecluse, Y., Plo, I., Dreyfus, F., Marzac, C., Casadevall, N., Lacombe, C., Romana, S., Dessen, P., Soulier, J., Viguie, F., Fontenay, M., Vainchenker, W., and Bernard, O. (2009). Mutation in TET2 in myeloid cancers. *N Engl J Med*, 360:2289–2301.
- Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56:52–64.
- Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V., Davies, R., Down, T. A., Haeffliger, C., Horton, R., Howe, K., Jackson, D. K., Kunde, J., Koenig, C., Liddle, J., Niblett, D., Otto, T., Pettett, R., Seemann, S., Thompson, C., West, T., Rogers, J., Olek, A., Berlin, K., and Beck, S. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature genetics*, 38(12):1378–1385.
- ElasticSearch (2010). ElasticSearch. <http://www.elasticsearch.org/>.
- ENCODE Common Cell Types (2010-2013). ENCODE common cell types. <http://genome.ucsc.edu/ENCODE/cellTypes.html>.
- EpiExplorer: supplementary information (2012). EpiExplorer: supplementary information.
- EpiGRAPH attribute documentation (2009). EpiGRAPH attribute documentation. <http://epigraph.mpi-inf.mpg.de/attributes/>.
- EpiGRAPH tutorial (2009). EpiGRAPH Tutorial. [http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH\\_tutorial.pdf](http://epigraph.mpi-inf.mpg.de/documentation/EpiGRAPH_tutorial.pdf).
- Epsztejn-Litman, S., Feldman, N., Abu-Remaileh, M., Shufaro, Y., Gerson, A., Ueda, J., Deplus, R., Fuks, F., Shinkai, Y., Cedar, H., and Bergman, Y. (2008). De novo DNA methylation promoted by g9a prevents reprogramming of embryonically silenced genes. *Nature structural & molecular biology*, 15(11):1176–1183.
- Ernst, J., Kheradpour, P., Mikkelsen, T., Shores, N., Ward, L., Epstein, C., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M., and Bernstein, B. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473:43–49.
- Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, 8(4):286–298.
- Fang, F., Fan, S., Zhang, X., and Zhang, M. Q. (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, 22(18):2204–2209.
- Feinberg, A. P. and Tycko, B. (2004). The history of cancer epigenetics. *Nature Reviews Cancer*, 4(2):143–153.
- Feuerbach, L. (2014). *Evolutionary Epigenomics - identifying functional genome elements by epigenetic footprints in the DNA*. Doctoral dissertation, Universität des Saarlandes, Saarbrücken.

- Feuerbach, L., Halachev, K., Assenov, Y., Müller, F., Bock, C., and Lengauer, T. (2012). Analyzing epigenome data in context of genome evolution and human diseases. In *Evolutionary Genomics*, pages 431–467. Springer.
- Ficz, G., Branco, M. R., Seisenberger, S., Santos, F., Krueger, F., Hore, T. A., Marques, C. J., Andrews, S., and Reik, W. (2011). Dynamic regulation of 5-hydroxymethylcytosine in mouse ES cells and during differentiation. *Nature*, 473(7347):398–402.
- Figuerola, M. E., Lugthart, S., Li, Y., Erpelinck-Verschueren, C., Deng, X., Christos, P. J., Schifano, E., Booth, J., van Putten, W., Skrabanek, L., Campagne, F., Mazumdar, M., Greally, J. M., Valk, P. J. M., Löwenberg, B., Delwel, R., and Melnick, A. (2010). DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. *Cancer cell*, 17(1):13–27.
- Fisher, R. A. (1922). On the Interpretation of  $\chi^2$  from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.
- Flicek, P., Aken, B., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T., Down, T., Dyer, S., Eyre, T., Fitzgerald, S., Fernandez-Banet, J., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Johnson, N., Jenkinson, A., Kahari, A., Keefe, D., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., and Megy, K. (2008). Ensembl 2008. *Nucleic Acids Res*, 36:D707–714.
- Flicek, P., Aken, B. L., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Gräf, S., Haider, S., Hammond, M., Howe, K., Jenkinson, A., Johnson, N., Kiliński, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Koscielny, G., Kulesha, E., Lawson, D., Longden, I., Massingham, T., McLaren, W., Megy, K., Overduin, B., Pritchard, B., Rios, D., Ruffier, M., Schuster, M., Slater, G., Smedley, D., Spudich, G., Tang, Y. A., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S. P., Zadissa, A., Birney, E., Cunningham, F., Dunham, I., Durbin, R., Fernández-Suarez, X. M., Herrero, J., Hubbard, T. J. P., Parker, A., Proctor, G., Smith, J., and Searle, S. M. J. (2010). Ensembl’s 10th year. *Nucleic Acids Research*, 38(Database issue):D557–562.
- Frommer, M., McDonald, L. E., Millar, D. S., Collis, C. M., Watt, F., Grigg, G. W., Molloy, P. L., and Paul, C. L. (1992). A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. *Proceedings of the National Academy of Sciences*, 89(5):1827–1831.
- Fujita, P. A., Rhead, B., Zweig, A. S., Hinrichs, A. S., Karolchik, D., Cline, M. S., Goldman, M., Barber, G. P., Clawson, H., Coelho, A., Diekhans, M., Dreszer, T. R., Giardine, B. M., Harte, R. A., Hillman-Jackson, J., Hsu, F., Kirkup, V., Kuhn, R. M., Learned, K., Li, C. H., Meyer, L. R., Pohl, A., Raney, B. J., Rosenbloom, K. R., Smith, K. E., Haussler, D., and Kent, W. J. (2011). The UCSC genome browser database: update 2011. *Nucleic Acids Research*, 39(Database issue):D876–882.
- Gardiner, E., Hunter, C., Packer, M., Palmer, D., and Willett, P. (2003). Sequence-dependent DNA structure: a database of octamer structural parameters. *J Mol Biol*, 332:1025–1035.

- Gardiner-Garden, M. and Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, 196(2):261–282.
- Gentleman, R., Carey, V., Bates, D., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J., and Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5:R80.
- Gentleman, R. and Lang, D. (2004). Statistical analyses and reproducible research. *Bioconductor Project Working Papers*. Paper 2.
- Giardine, B., Riemer, C., Hardison, R., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J., Miller, W., Kent, W., and Nekrutenko, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. *Genome Res*, 15:1451–1455.
- Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11:R86.
- Google Chart Tools (2010-2013). Google Chart Tools. <http://code.google.com/apis/chart/>.
- Greenbaum, J., Pang, B., and Tullius, T. (2007). Construction of a genome-scale structural map at single-nucleotide resolution. *Genome Res*, 17:947–953.
- Gu, H., Bock, C., Mikkelsen, T. S., Jäger, N., Smith, Z. D., Tomazou, E., Gnirke, A., Lander, E. S., and Meissner, A. (2010). Genome-scale DNA methylation mapping of clinical samples at single-nucleotide resolution. *Nature methods*, 7(2):133–136.
- Halachev, K. (2006). Epigraph\*regression: A toolkit for (epi-)genomic correlation analysis and prediction of quantitative attributes. Master’s thesis, Universität des Saarlandes.
- Halachev, K., Bast, H., Albrecht, F., Lengauer, T., and Bock, C. (2012). EpiExplorer: live exploration and global analysis of large epigenomic datasets. *Genome Biology*, 13(10):R96.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The WEKA data mining software: an update. *Special Interest Group on Knowledge Discovery and Data Mining Explorer Newsletter*, 11(1):10–18.
- Hamosh, A., Scott, A., Amberger, J., Bocchini, C., and McKusick, V. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*, 33:D514–517.
- Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*, volume 2. Springer.
- Hawkins, R., Hon, G., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11:476–486.

- HBase (2008). HBase. <http://hbase.apache.org/>.
- Hearst, M. (2009). *Search User Interfaces*. Cambridge, New York: Cambridge University Press.
- Heintzman, N., Hon, G., Hawkins, R., Kheradpour, P., Stark, A., Harp, L., Ye, Z., Lee, L., Stuart, R., Ching, C., Ching, K., Antosiewicz-Bourget, J., Liu, H., Zhang, X., Green, R., Lobanenko, V., Stewart, R., Thomson, J., Crawford, G., Kellis, M., and Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, 459:108–112.
- Holliday, R. and Grigg, G. (1993). Dna methylation and mutation. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 285(1):61–67.
- Huang, D., Sherman, B., Tan, Q., Collins, J., Alvord, W., Roayaei, J., Stephens, R., Baseler, M., Lane, H., and Lempicki, R. (2007). The DAVID gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 8:R183.
- Huang, Y., Pastor, W. A., Shen, Y., Tahiliani, M., Liu, D. R., and Rao, A. (2010). The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. *PLoS One*, 5(1):e8888.
- Huarte, M., Guttman, M., Feldser, D., Garber, M., Koziol, M., Kenzelmann-Broz, D., Khalil, A., Zuk, O., Amit, I., Rabani, M., Attardi, L., Regev, A., Lander, E., Jacks, T., and Rinn, J. (2010). A large intergenic noncoding RNA induced by p53 mediates global gene repression in the p53 response. *Cell*, 142:409–419.
- Hudson, T. J., Anderson, W., Aretz, A., Barker, A. D., Bell, C., Bernabé, R. R., Bhan, M., Calvo, F., Eerola, I., Gerhard, D. S., et al. (2010). International network of cancer genome projects. *Nature*, 464(7291):993–998.
- Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., and Oinn, T. (2006). Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web Server):W729–W732.
- Human Epigenome Atlas (2010-2013). Human Epigenome Atlas. <http://www.epigenomeatlas.org/>.
- Illingworth, R., Kerr, A., DeSousa, D., Jørgensen, H., Ellis, P., Stalker, J., Jackson, D., Clee, C., Plumb, R., Rogers, J., Humphray, S., Cox, T., Langford, C., and Bird, A. (2008). A novel cpg island set identifies tissue-specific methylation at developmental gene loci. *PLoS biology*, 6(1):e22.
- Illingworth, R. S., Gruenewald-Schneider, U., Webb, S., Kerr, A. R., James, K. D., Turner, D. J., Smith, C., Harrison, D. J., Andrews, R., and Bird, A. P. (2010). Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS genetics*, 6(9):e1001134.
- Illustration of the X-GRAF File Format (2009). Illustration of the X-GRAF File Format. [http://epigraph.mpi-inf.mpg.de/documentation/X-GRAF\\_Format\\_Illustration.pdf](http://epigraph.mpi-inf.mpg.de/documentation/X-GRAF_Format_Illustration.pdf).

- Ioshikhes, I. P. and Zhang, M. Q. (2000). Large-scale human promoter mapping using CpG islands. *Nature genetics*, 26(1):61–63.
- Ito, S., D'Alessio, A. C., Taranova, O. V., Hong, K., Sowers, L. C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature*, 466(7310):1129–1133.
- Jackson, M., Krassowska, A., Gilbert, N., Chevassut, T., Forrester, L., Ansell, J., and Ramsahoye, B. (2004). Severe global DNA hypomethylation blocks differentiation and induces histone hyperacetylation in embryonic stem cells. *Molecular and cellular biology*, 24(20):8862–8871.
- Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nature genetics*, 33:245–254.
- Java (2009). Java. <http://www.java.com/>.
- Java Architecture for XML Binding (2009). Java Architecture for XML Binding. <https://jaxb.dev.java.net/>.
- Jiang, C., Han, L., Su, B., Li, W.-H., and Zhao, Z. (2007). Features and trend of loss of promoter-associated CpG islands in the human and mouse genomes. *Molecular biology and evolution*, 24(9):1991–2000.
- Jones, P. A. and Liang, G. (2009). Rethinking how dna methylation patterns are maintained. *Nature Reviews Genetics*, 10(11):805–811.
- jQuery (2010-2013). jQuery. <http://jquery.org/>.
- Karolchik, D., Hinrichs, A., Furey, T., Roskin, K., Sugnet, C., Haussler, D., and Kent, W. (2004). The UCSC table browser data retrieval tool. *Nucleic Acids Res*, 32:D493–496.
- Karolchik, D., Hinrichs, A. S., and Kent, W. J. (2011). The UCSC genome browser. *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]*, Chapter 18:Unit18.6.
- Karolchik, D., Kuhn, R., Baertsch, R., Barber, G., Clawson, H., Diekhans, M., Giardine, B., Harte, R., Hinrichs, A., Hsu, F., Kober, K., Miller, W., Pedersen, J., Pohl, A., Raney, B., Rhead, B., Rosenbloom, K., Smith, K., Stanke, M., Thakkapallayil, A., Trumbower, H., Wang, T., Zweig, A., Haussler, D., and Kent, W. (2008). The UCSC genome browser database: 2008 update. *Nucleic Acids Res*, 36:D773–779.
- Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*, 14:160–169.
- Keshet, I., Lieman-Hurwitz, J., and Cedar, H. (1986). DNA methylation affects the formation of active chromatin. *Cell*, 44(4):535–543.

- Kinsella, R., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., Almeida-King, J., Staines, D., Derwent, P., Kerhornou, A., Kersey, P., and Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, 2011:bar030.
- Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, 128(4):693–705.
- Kriaucionis, S. and Heintz, N. (2009). The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science*, 324:929–930.
- Krzywinski, M. I., Schein, J. E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research*.
- Ku, M., Koche, R. P., Rheinbay, E., Mendenhall, E. M., Endoh, M., Mikkelsen, T. S., Presser, A., Nusbaum, C., Xie, X., Chi, A. S., Adli, M., Kasif, S., Ptaszek, L. M., Cowan, C. A., Lander, E. S., Koseki, H., and Bernstein, B. E. (2008). Genomewide Analysis of PRC1 and PRC2 Occupancy Identifies Two Classes of Bivalent Domains. *PLoS Genet*, 4(10):e1000242+.
- Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics*, 11(3):191–203.
- Laurent, L., Wong, E., Li, G., Huynh, T., Tsigos, A., Ong, C. T., Low, H. M., Sung, K. W. K., Rigoutsos, I., Loring, J., and Wei, C.-L. (2010). Dynamic changes in the human methylome during differentiation. *Genome research*, 20(3):320–331.
- Laurent, S. S., Johnston, J., and Dumbill, E. (2001). *Programming web services with XML-RPC*. O’reilly.
- Leonhardt, H., Page, A. W., Weier, H.-U., and Bestor, T. H. (1992). A targeting sequence directs DNA methyltransferase to sites of DNA replication in mammalian nuclei. *Cell*, 71(5):865–873.
- Li, E., Beard, C., and Jaenisch, R. (1993). Role for DNA methylation in genomic imprinting. *Nature*, 366(6453):362–365.
- Li, E., Bestor, T. H., and Jaenisch, R. (1992). Targeted mutation of the dna methyltransferase gene results in embryonic lethality. *Cell*, 69(6):915–926.
- Mann, H. B. and Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge.
- Marchionini, G. (2006). Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46.

- Matsuo, K., Clay, O., Takahashi, T., Silke, J., and Schaffner, W. (1993). Evidence for erosion of mouse CpG islands during mammalian evolution. *Somatic cell and molecular genetics*, 19(6):543–555.
- Maunakea, A. K., Nagarajan, R. P., Bilenky, M., Ballinger, T. J., D’Souza, C., Fouse, S. D., Johnson, B. E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V. M., Rowitch, D. H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S. J., Haussler, D., Marra, M. A., Hirst, M., Wang, T., and Costello, J. F. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature*, 466(7303):253–257.
- McLean, C. Y., Bristor, D., Hiller, M., Clarke, S. L., Schaar, B. T., Lowe, C. B., Wenger, A. M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5):495–501.
- Meissner, A., Mikkelsen, T. S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B. E., Nusbaum, C., Jaffe, D. B., Gnirke, A., Jaenisch, R., and Lander, E. S. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, 454(7205):766–770.
- Mendenhall, E. M., Koche, R. P., Truong, T., Zhou, V. W., Issac, B., Chi, A. S., Ku, M., and Bernstein, B. E. (2010). GC-Rich Sequence Elements Recruit PRC2 in Mammalian ES Cells. *PLoS Genet*, 6(12):e1001244+.
- Mikkelsen, T., Xu, Z., Zhang, X., Wang, L., Gimble, J., Lander, E., and Rosen, E. (2010). Comparative epigenomic analysis of murine and human adipogenesis. *Cell*, 143:156–169.
- Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.-K. K., Koche, R. P., Lee, W., Mendenhall, E., O’Donovan, A., Presser, A., Russ, C., Xie, X., Meissner, A., Wernig, M., Jaenisch, R., Nusbaum, C., Lander, E. S., and Bernstein, B. E. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153):553–560.
- Mitchell, P. and Tjian, R. (1989). Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science*, 245:371–378.
- Mohn, F., Weber, M., Rebhan, M., Roloff, T. C., Richter, J., Stadler, M. B., Bibel, M., and Schübeler, D. (2008). Lineage-specific polycomb targets and de novo DNA methylation define restriction and potential of neuronal progenitors. *Molecular cell*, 30(6):755–766.
- Mohr, F., Dohner, K., Buske, C., and Rawat, V. (2011). TET genes: new players in DNA demethylation and important determinants for stemness. *Exp Hematol*, 39:272–281.
- MongoDB (2009). MongoDB. <http://www.mongodb.com/>.
- Mouse ENCODE Consortium, Stamatoyannopoulos, J. A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D. M., Groudine, M., Bender, M., Kaul, R., Canfield, T., Giste, E., Johnson, A., Zhang, M., Balasundaram, G., Byron, R., Roach, V., Sabo, P. J., Sandstrom, R., Stehling, A. S., Thurman, R. E., Weissman, S. M., Cayting, P., Hariharan, M., Lian, J., Cheng, Y., Landt, S. G., Ma, Z., Wold, B. J., Dekker, J.,

- Crawford, G. E., Keller, C. A., Wu, W., Morrissey, C., Kumar, S. A., Mishra, T., Jain, D., Byrska-Bishop, M., Blankenberg, D., Lajoie, B. R., Jain, G., Sanyal, A., Chen, K.-B. B., Denas, O., Taylor, J., Blobel, G. A., Weiss, M. J., Pimkin, M., Deng, W., Marinov, G. K., Williams, B. A., Fisher-Aylor, K. I., Desalvo, G., Kiralusha, A., Trout, D., Amrhein, H., Mortazavi, A., Edsall, L., McCleary, D., Kuan, S., Shen, Y., Yue, F., Ye, Z., Davis, C. A., Zaleski, C., Jha, S., Xue, C., Dobin, A., Lin, W., Fastuca, M., Wang, H., Guigo, R., Djebali, S., Lagarde, J., Ryba, T., Sasaki, T., Malladi, V. S., Cline, M. S., Kirkup, V. M., Learned, K., Rosenbloom, K. R., Kent, W. J., Feingold, E. A., Good, P. J., Pazin, M., Lowdon, R. F., and Adams, L. B. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome biology*, 13(8):418.
- Munzel, M., Globisch, D., and Carell, T. (2011). 5-hydroxymethylcytosine, the sixth base of the genome. *Angew Chem Int Ed Engl*, 50:6460–6468.
- Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N., Ahfeldt, T., Sachs, K., Li, X., Li, H., Kuperwasser, N., Ruda, V., Pirruccello, J., Muchmore, B., Prokunina-Olsson, L., Hall, J., Schadt, E., Morales, C., Lund-Katz, S., Phillips, M., Wong, J., Cantley, W., Racie, T., Ejebe, K., Orho-Melander, M., Melander, O., Koteliensky, V., Fitzgerald, K., Krauss, R., Cowan, C., Kathiresan, S., and Rader, D. (2010). From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466:714–719.
- Myers, R., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R., Bernstein, B., Gingeras, T., Kent, W., Birney, E., Wold, B., and Crawford, G. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, 9:e1001046.
- MySQL Database (1995). MySQL Database. <http://www.mysql.com/>.
- Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., and Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. *Bioinformatics (Oxford, England)*, 28(14):1919–1920.
- Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak, R. G. W., Hoadley, K. A., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding, L., Wilson, R. K., Van Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., and Aldape, K. (2010). Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell*, 17(5):510–522.
- Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M. R., Senger, M., Stevens, R., Wipat, A., and Wroe, C. (2006). Taverna: lessons in creating a workflow environment for the life sciences: Research articles. *Concurr. Comput. : Pract. Exper.*, 18(10):1067–1100.
- Okano, M., Bell, D. W., Haber, D. A., and Li, E. (1999). DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell*, 99(3):247–257.



- Oracle Database (2009). Oracle Database. <http://www.oracle.com/database/>.
- Oracle XML DB (2009). Oracle XML DB. <http://www.oracle.com/technology/tech/xml/xmlldb/index.html>.
- Orkin, S. (1990). Globin gene regulation and switching: circa 1990. *Cell*, 63:665–672.
- Pauler, F. M., Sloane, M. A., Huang, R., Regha, K., Koerner, M. V., Tamir, I., Sommer, A., Aszodi, A., Jenuwein, T., and Barlow, D. P. (2009). H3K27me3 forms BLOCs over silent genes and intergenic regions and specifies a histone banding pattern on a mouse autosomal chromosome. *Genome research*, 19(2):221–233.
- Popp, C., Dean, W., Feng, S., Cokus, S. J., Andrews, S., Pellegrini, M., Jacobsen, S. E., and Reik, W. (2010). Genome-wide erasure of dna methylation in mouse primordial germ cells is affected by aid deficiency. *Nature*, 463(7284):1101–1105.
- Python Programming Language (2009). Python Programming Language. <http://www.python.org/>.
- Qiu, J. (2006). Epigenetics: unfinished symphony. *Nature*, 441(7090):143–145.
- Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.
- R Project for Statistical Computing (2009). R Project for Statistical Computing. <http://www.r-project.org/>.
- Rakyan, V. K., Down, T. A., Thorne, N. P., Flicek, P., Kulesha, E., Gräf, S., Tomazou, E. M., Bäckdahl, L., Johnson, N., Herberth, M., Howe, K. L., Jackson, D. K., Miretti, M. M., Fiegler, H., Marioni, J. C., Birney, E., Hubbard, T. J. P., Carter, N. P., Tavaré, S., and Beck, S. (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome research*, 18(9):1518–1529.
- Raney, B., Cline, M., Rosenbloom, K., Dreszer, T., Learned, K., Barber, G., Meyer, L., Sloan, C., Malladi, V., Roskin, K., Suh, B., Hinrichs, A., Clawson, H., Zweig, A., Kirkup, V., Fujita, P., Rhead, B., Smith, K., Pohl, A., Kuhn, R., Karolchik, D., Haussler, D., and Kent, W. (2011). ENCODE whole-genome data in the UCSC genome browser (2011 update). *Nucleic Acids Res*, 39:D871–875.
- Redis (2009). Redis. <http://redis.io/>.
- Reich, M., Liefeld, J., Thorvaldsdottir, H., Ocana, M., Polk, E., Jang, D., and Mesirov, J. (2012). Genomespace: An environment for frictionless bioinformatics. In *Proceedings of the 103rd Annual Meeting of the American Association for Cancer Research*, volume 72, page 3966.
- Reik, W. (2007). Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143):425–432.

- Reik, W., Dean, W., and Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, 293(5532):1089–1093.
- Reik, W. and Walter, J. (2001). Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32.
- Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. *Nature biotechnology*, 29(1):24–26.
- Rollins, R., Haghighi, F., Edwards, J., Das, R., Zhang, M., Ju, J., and Bestor, T. (2006). Large-scale structure of genomic methylation patterns. *Genome Res*, 16:157–163.
- Russo, V. E. A., Martienssen, R. A., and Riggs, A. D. (1996). *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press.
- Sandve, G., Gundersen, S., Rydbeck, H., Glad, I., Holden, L., Holden, M., Liestol, K., Clancy, T., Ferkingstad, E., Johansen, M., Nygaard, V., Tostesen, E., Frigessi, A., and Hovig, E. (2010). The Genomic HyperBrowser: inferential genomics at the sequence level. *Genome Biol*, 11:R121.
- Satterlee, J., Schubeler, D., and Ng, H. (2010). Tackling the epigenome: challenges and opportunities for collaboration. *Nat Biotechnol*, 28:1039–1044.
- Schuster, S. C. (2007). Next-generation sequencing transforms today’s biology. *Nature*, 200(8).
- Siepel, A., Bejerano, G., Pedersen, J., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L., Richards, S., Weinstock, G., Wilson, R., Gibbs, R., Kent, W., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*, 15:1034–1050.
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCr: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.
- Smallwood, S. A., Tomizawa, S.-i., Krueger, F., Ruf, N., Carli, N., Segonds-Pichon, A., Sato, S., Hata, K., Andrews, S. R., and Kelsey, G. (2011). Dynamic CpG island methylation landscape in oocytes and preimplantation embryos. *Nature genetics*, 43(8):811–814.
- Smit, A., Hubley, R., and Green, P. (1996-2010). RepeatMasker Open-3.0. <http://www.repeatmasker.org/>.
- Smith, Z. D. and Meissner, A. (2013). DNA methylation: roles in mammalian development. *Nature Reviews Genetics*.
- Song, F., Smith, J. F., Kimura, M. T., Morrow, A. D., Matsuyama, T., Nagase, H., and Held, W. A. (2005). Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 102(9):3336–3341.

- Soria, G., Polo, S. E., and Almouzni, G. (2012). Prime, repair, restore: the active role of chromatin in the DNA damage response. *Molecular Cell*, 46(6):722–734.
- Stadler, M., Murr, R., Burger, L., Ivanek, R., Lienert, F., Scholer, A., Wirbelauer, C., Oakeley, E., Gaidatzis, D., Tiwari, V., and Schubeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, 480:490–495.
- Sterner, D. E. and Berger, S. L. (2000). Acetylation of histones and transcription-related factors. *Microbiology and Molecular Biology Reviews*, 64(2):435–459.
- Straussman, R., Nejman, D., Roberts, D., Steinfeld, I., Blum, B., Benvenisty, N., Simon, I., Yakhini, Z., and Cedar, H. (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nature structural & molecular biology*, 16(5):564–571.
- Stroud, H., Feng, S., Morey Kinney, S., Pradhan, S., and Jacobsen, S. (2011). 5-hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol*, 12:R54.
- Su, A., Wiltshire, T., Batalov, S., Lapp, H., Ching, K., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., Cooke, M., Walker, J., and Hogenesch, J. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA*, 101:6062–6067.
- Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J. (2007). GSEA-P: a desktop application for gene set enrichment analysis. *Bioinformatics*, 23:3251–3253.
- Szulwach, K., Li, X., Li, Y., Song, C., Han, J., Kim, S., Namburi, S., Hermetz, K., Kim, J., Rudd, M., Yoon, Y., Ren, B., He, C., and Jin, P. (2011). Integrating 5-hydroxymethylcytosine into the epigenomic landscape of human embryonic stem cells. *PLoS Genet*, 7:e1002154.
- Tahiliani, M., Koh, K., Shen, Y., Pastor, W., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L., Liu, D., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science*, 324:930–935.
- Takai, D. and Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*, 99(6):3740–3745.
- Tološi, L., Theißen, J., Halachev, K., Hero, B., Berthold, F., and Lengauer, T. (2013). A method for finding consensus breakpoints in the cancer genome from copy number data. *Bioinformatics*.
- Tucker, K., Beard, C., Dausmann, J., Jackson-Grusby, L., Laird, P. W., Lei, H., Li, E., and Jaenisch, R. (1996). Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes & development*, 10(8):1008–1020.
- Tunkelang, D. (2009). *Faceted Search*. San Rafael, CA: Morgan & Claypool Publishers.

- UCSC Genome Browser BED format documentation (2011). UCSC genome browser BED format documentation. <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>.
- van Steensel, B. (2005). Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat Genet*, 37(Suppl):S18–24.
- Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E. M., and Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, 457(7231):854–858.
- Waddington, C. (1942). The epigenotype. *Endeavour*, 1:18–20.
- Weber, M., Davies, J. J., Wittig, D., Oakeley, E. J., Haase, M., Lam, W. L., and Schuebeler, D. (2005). Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature genetics*, 37(8):853–862.
- Weber, M., Hellmann, I., Stadler, M. B., Ramos, L., Paabo, S., Rebhan, M., and Schubeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nat Genet*, 39(4):457–466.
- Wu, H., D'Alessio, A. C., Ito, S., Xia, K., Wang, Z., Cui, K., Zhao, K., Sun, Y. E., and Zhang, Y. (2011). Dual functions of Tet1 in transcriptional regulation in mouse embryonic stem cells. *Nature*, 473(7347):389–393.
- Wu, S. C. and Zhang, Y. (2010). Active DNA demethylation: many roads lead to Rome. *Nature Reviews Molecular Cell Biology*, 11(9):607–620.
- X-GRAF XML Format Documentation (2009). X-GRAF XML Format Documentation. <http://epigraph.mpi-inf.mpg.de/xml/>.
- Xhemalce, B., Dawson, M. A., and Bannister, A. J. (2011). Histone modifications. *Encyclopedia of Molecular Cell Biology and Molecular Medicine*.
- Yamada, Y., Shirakawa, T., Taylor, T. D., Okamura, K., Soejima, H., Uchiyama, M., Iwasaka, T., Mukai, T., Muramoto, K., Sakaki, Y., and Ito, T. (2006). A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 11q: comparison with chromosome 21q. *DNA sequence : the journal of DNA sequencing and mapping*, 17(4):300–306.
- Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., Mukai, T., Sakaki, Y., and Ito, T. (2004). A Comprehensive Analysis of Allelic Methylation Status of CpG Islands on Human Chromosome 21q. *Genome Research*, 14(2):247–266.
- Yi, J. M., Dhir, M., Van Neste, L., Downing, S. R., Jeschke, J., Glöckner, S. C., de Freitas Calmon, M., Hooker, C. M., Funes, J. M., Boshoff, C., Smits, K. M., van Engeland, M., Weijenberg, M. P., Iacobuzio-Donahue, C. A., Herman, J. G., Schuebel, K. E., Baylin, S. B., and Ahuja, N. (2011). Genomic and epigenomic integration identifies a prognostic signature in colon cancer. *Clinical Cancer Research*, 17(6):1535–1545.

- Zentner, G. E. and Henikoff, S. (2013). Regulation of nucleosome dynamics by histone modifications. *Nature structural & molecular biology*, 20(3):259–266.
- Zentner, G. E., Tesar, P. J., and Scacheri, P. C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome research*, 21(8):1273–1283.
- Zhang, Y. and Reinberg, D. (2001). Transcription regulation by histone methylation: interplay between different covalent modifications of the core histone tails. *Genes & development*, 15(18):2343–2360.
- Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E. A., Koebbe, B. C., Nielsen, C., Hirst, M., Farnham, P., Kuhn, R. M., Zhu, J., Smirnov, I., Kent, W. J., Haussler, D., Madden, P. A. F., Costello, J. F., and Wang, T. (2011). The Human Epigenome Browser at Washington University. *Nat Meth*, 8(12):989–990.
- Zobel, J., Moffat, A., and Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing. *ACM Transactions on Database Systems (TODS)*, 23(4):453–490.